DOCUMENT RESUME

ED 148 894 TM 006 834

AUTHOR Gephart, William J.; And Others
TITLE The Evaluation of Teaching--NSPER: 74. Occasional Paper 18.
INSTITUTION Phi Delta Kappa, Bloomington, Ind. Center on Evaluation and Research.
PUB DATE 74
NOTE 104p.
AVAILABLE FROM Phi Delta Kappa, Eighth and Union, Box 789, Bloomington, Indiana 47401 ($2.00)

EDRS PRICE MF-$0.83 HC-$6.01 Plus Postage.
DESCRIPTORS College Teachers; *Conference Reports; Course Evaluation; Educational Assessment; Elementary School Teachers; Elementary Secondary Education; *Evaluation Criteria; *Evaluation Methods; Measurement Goals; Post Secondary Education; Rating Scales; Secondary School Teachers; *Teacher Evaluation; Teacher Improvement; Teaching Procedures; *Teaching Quality; *Teaching Styles; Workshops

ABSTRACT

The 1974 National Symposia for Professors of Educational Research (NSPER), which concerned the evaluation of teaching, are summarized. Each session consisted of three components: (1) a Teacher Appraisal for Improvement workshop, focusing upon methods for appraising teaching, (2) a discussion of the criteria which should be used to evaluate teaching, and (3) small group activities designed to: assess the utility of these concepts and procedures, and to plan the implementation of these and other methods of evaluation in specific contexts (i.e., elementary, secondary, and post secondary educational institutions). Three related purposes for evaluation were identified: to produce information for the improvement of an individual teacher's instructional skills; for administrative decisions on hiring, firing, promotion and tenure; and for the reallocation of resources. The purposes of teaching, and various styles of teaching are discussed. Different methods of teacher evaluation are presented; a brief review of relevant research and guidelines for constructing teacher rating scales are included. A brief evaluation of the symposia is appended. (MV)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

PHI DELTA KAPPA

EIGHTH AND UNION BOX 789
BLOOMINGTON INDIANA 47401

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED148894

PRESENTS AN

OCCASIONAL PAPER

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

PHI DELTA KAPPA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM 006 834

FROM

CENTER ON EVALUATION DEVELOPMENT AND RESEARCH

CEDR

OCCASIONAL PAPER 18
THE EVALUATION OF TEACHING – NSPER: 74

by

William J. Gephart
Phi Delta Kappa

Robert B. Ingle
University of Wisconsin–Milwaukee

Gary Saretsky
Phi Delta Kappa

2A

AN OVERVIEW OF THE REPORT

This report summarizes the three (Denver, Seattle, and Atlanta) meetings of the National Symposium for Professors of Educational Research (NSPER: 74) whose topic was evaluating teaching.

The report is prefaced with a description of the origins, purpose, and evolution of NSPER. An overview of the three NSPER sessions follows which summarizes the major points that surfaced at the three meetings. The question of what aspects of teaching should an evaluation focus upon was addressed by Christine McGuire, Harry Broudy, and Irvin Lehman respectively at the Denver, Seattle, and Atlanta sessions. These presentations are reproduced beginning on page 7. A major component of the three sessions was the Center for the Study of Evaluation workshop *Teacher Appraisal For Improvement* (TAFI). A description of the workshop and report of participants' reactions to the workshop are presented on pages 93 through 102.

THE ORIGINS OF THE NATIONAL SYMPOSIUM FOR PROFESSORS OF EDUCATIONAL RESEARCH (NSPER)

Concern for improvement of instruction about research and related processes in the mid-sixties was the impetus for NSPER. At that time many people were very critical of research training programs. Simultaneously, more and more universities were developing research training programs. People assigned to teach in these programs frequently felt ill-prepared for the assignment. The National Symposium for Professors of Educational Research was proposed to Phi Delta Kappa's Board of Directors as one effort to improve research instruction. Two general objectives for NSPER were stated: 1) to provide an opportunity for people who teach about the research and research related processes to exchange information about problems, materials, and techniques in their instructional assignments; and 2) to help those individuals obtain new information from leaders in the field.

During the past decade, significant advances were made toward the development of a technology of evaluation. Since many of the techniques utilized in the evaluation process were also used in the research process, responsibility for teaching about educational evaluation was assigned to or assumed by professors of educational research. As a consequence, the focus of NSPER has expanded to include educational evaluation as an appropriate topic for its annual symposia. NSPER has also undergone an expansion in terms of types of participants. Originally the participants were primarily professors of research. This has expanded in the last two years to include individuals who have supervisory or leadership roles for research or evaluation work carried out in educational institutions.

The first session of NSPER, held in 1967 at the University of Maryland, focused on alternative approaches for teaching the four basic research areas: introduction to research, measurement, statistics, and advanced design. The 56 participants came from universities in more than 30 states, Canada, and England. Their reactions to the opportunity to share information about content to be taught, teaching techniques, and resources resulted in strong recommendations that the symposium be repeated. Subsequent sessions have focused on the topics shown below.

NSPER: 68 – Instructional Materials Development for Research Instruction (University of Colorado)

NSPER: 69 – Research Techniques from Selected Disciplines that are Possibly Applicable to Education (University of Chicago)

NSPER: 70 – The Introductory Research Course: Content, Technique, and Resources (Washington University, St. Louis)

NSPER 71 – Graduate Measurement Instruction (Held in three locations: University of Wisconsin–Milwaukee, University of Florida, and University of Arizona)

NSPER 72 – Similarities and Differences in the Research and Evaluation Processes (Held in three locations: Michigan State University, San Francisco, and University of North Carolina)

NSPER 73 – Teaching about Evaluation (Held in four loca-

tions: University of Wisconsin–Milwaukee, SUNY–Albany, UCLA, and Chicago)

The focus of NSPER: 73 was upon "Teaching About Evaluation." Different instructional materials were utilized and informally critiqued at the four different sessions. At the UCLA session, participants had the opportunity to work through portions of a pilot version of the Center for the Study of Evaluation's (CSE) Teacher Appraisal for Improvement (TAFI) workshop. Subsequent interaction among NSPER: 73 participants and consultation with Phi Delta Kappa's Research Advisory Committee led to the decision of focusing NSPER: 74 upon the evaluation of teaching, and using the TAFI workshop as part of the symposium.

AN OVERVIEW OF THE NSPER: 74 SESSION ON EVALUATING TEACHING

Continuing an effort to make NSPER sessions more accessible to participants, the NSPER: 74 sessions were held in three different sites: Denver, Seattle, and Atlanta. Each two and one-half day session was made up of three components.

1. A one-day TAFI workshop focusing upon procedures and techniques useful in appraising teaching.
2. A guest presentation addressing the question: What aspects of teaching should an evaluation focus upon?
3. Small group activities designed to
   a. assess the utility of the concepts, techniques, and procedures presented via 1 and 2 above, and
   b. if appropriate, to develop a plan for the use of these and other concepts, techniques, and procedures to achieve the various purposes of evaluation in specific contexts (i.e., elementary, secondary, and post-secondary educational institutions).

THE APPRAISAL OF TEACHING: PURPOSE, FOCUS, AND MEANS

Three related questions appeared to be addressed through the three NSPER sessions: 1) What are the purposes of evaluating

4

teaching? 2) What entities or phenomenon should an evaluation focus upon? and 3) What are the methods, procedures, and techniques that may prove useful in undertaking the evaluation of teaching? These questions are addressed at length in the papers on pages 7 through 92. The following, however, are the major points that surfaced in the three meetings.

1. Purposes of Evaluation

At least three related but different purposes can be identified:

a. Evaluation which produces information for the improvement of an individual teacher's instructional skills.
b. Evaluation which produces information for administrative decisions on hiring, firing, promotion, and tenure.
c. Evaluation which produces information for the reallocation of resources necessary for the improvement of the teaching-learning process.

In light of the differing purposes of evaluation, it was deemed necessary that the design of any procedures for evaluating teaching *should start with* a clear determination of the purpose of the evaluation work to be undertaken. Furthermore, it appeared that the accomplishment of these differing purposes would require the creation of separate (but interfaced) evaluation plans. (This seems necessary due to the amount of threat and anxiety currently felt by teachers, the need for self-disclosure and analysis inherent in the improvement of instruction, and the need for objectivity, documentation, and visibility in administrative decision making.)

2. Foci of Evaluation: Modes, Processes, and Product

Three modes of teaching can be identified which make up teaching behavior. All three of them are present all of the time. BUT, at any instance one of these modes is dominant or primary and the other two are less dominant or secondary. The three modes are:

a. Didactic teaching – actions by the teacher designed to

help learners master facts and principles.

b. Heuristic teaching – actions by the teacher designed to help learners develop problem-solving skills and strategies.

c. Philetic teaching – actions by the teacher designed to help learners develop interpersonal skills and relationships.

Dr. Harry Broudy, who proposed this three-way view of teaching, asserted that in the evaluation of teaching we must first know what mode of teaching is dominant, and second, use criteria appropriate to that dominant mode in the evaluation of that teaching-learning enterprise.

Evaluation of teaching has BOTH process and product components. That is, we need to look at both what is done by the teacher and what is accomplished or learned by the student. This latter point is based on the assumption that the learner has the ultimate control over how his or her energy will be expended. It is possible for a teacher to do all of the generally accepted teacher activities and for a student to decide not to exert the energy needed to master the lesson. In such cases the teacher ought not to be faulted. This is not different from the physician who follows accepted medical practice but whose patient succumbs to the illness anyway.

3. Means of Evaluation

There are eight different means or approaches identified in the materials developed by the UCLA Center for the Study of Evaluation and used in each of the three NSPER: 74 sessions. *Each of them has advantages and disadvantages, strengths, and weaknesses.* Those eight are:

a. Unstructured Observations and Ratings (Subjective and idiosyncratic observation of teaching by administrators, peers, or students).

b. Systematic Observations and Ratings (Use of prespecified procedures for observing teaching and recording descriptions of teaching behaviors used by administrators, peers, or students. Flanders Interaction Analysis, teacher skill checklists, etc., are examples).

c. Teacher Skill Testing (The presentation of a teaching-learning situation in writing or by oral or videotape and asking teachers to indicate what should be done).
d. Omnibus Measures of Student Growth (The use of standardized tests to determine amount of student growth related to general educational goals and thus to infer the quality of teaching).
e. Objectives-based Measures of Student Growth (Testing students on specific objectives concentrated on during periods ranging from several weeks to a semester).
f. Very Short Measures of Student Growth (Testing students on the acquisition of facts or skills taught in a brief instructional period).
g. Teacher Performance Testing Systems (Structuring a teaching-learning situation concentrating on a single objective and observing teaching and learning related to that objective).
h. Analysis of Typical Student Growth (The use of regression analysis to identify effective teachers and subsequently the techniques used by teachers whose students perform at or above predicted achievement levels).

The following caveats were made regarding the above techniques: Because each of these approaches has weaknesses, the reliance on any one of them as *the approach* to evaluating teaching is not likely to be adequate. Furthermore, the selection and use of some combination of these and/or approaches should be based on the evaluation purpose in a given setting, and upon the needs and resources in that setting. In other words, the design of any approach to evaluating teaching should reflect and be responsive to the needs, resources, goals, and requirements of the local situation. In addition, it should be recognized that an evaluation plan that works well for assessing the quality of teaching in one context may not be successful in a different setting. Finally, most of these methodological approaches can and should be considered from two vantage points: self evaluation and evaluation by individuals other than the teacher being evaluated.

THREE MODES OF TEACHING AND THEIR EVALUATION

H. S. Broudy

The attempts to evaluate teaching have suffered less from a lack of zeal and intelligence than from an oversupply of theories about teaching and learning. Deciding which features of teaching to evaluate, therefore, is no less a problem than choosing the methods for judging them. In exasperation or desperation some of our researchers have decreed that only those aspects of teaching they can measure shall count as teaching. Until one analysis drives all or nearly all of its competitors from the textbooks, we shall have to cope with as many analyses, criteria, and methods of evaluation as can muster the support of an articulate coterie. Teaching as behavior modification with various modifications; teaching as carrying out an instructional program; teaching as classroom management; teaching as social interaction; teaching as socio-drama; teaching as psychological therapy (or welfare); teaching as the mystical communication of soul to soul; teaching as the nurturance of unfolding – these and any number of others have coteries of various sizes and influence. The allegiance of the bulk of the researchers is to some variant of B. F. Skinner's operant conditioning, whereas teachers and administrators, not to speak of the common run of mankind, resent what they regard as Skinner's reduction of the teacher-learner encounter to a form of psychic manipulation.

My own contribution to this plethora of wisdom about teaching differs from most of the others in at least two respects: first, it is much more simple-minded than most of them, and, second, it commands the support of no coterie that I have been able to identify.

I propose to conceive of the teaching act as being of three distinguishable types: didactics, heuristics, and philetics, and although all three, or two out of the three, may be operating together, at any given moment, only one of these is dominant. It is in the focus of the teacher's attention while the other or others recede into the background and operate tacitly or implicitly. The teaching act is composed of 1) an explicit manifest pattern of

activity with a distinctive style or structure and 2) components that one teaches with, but which are not at the center of the teacher's conscious efforts.*

Of these components, some can be analyzed and formulated as procedures. For example, certain aspects of classroom management could be so formulated, but some like confidence, control, personal charm, and attitudes toward pupils may not be. Each mode of teaching has a classroom atmosphere that is appropriate to it and a pattern of interpersonal relationships that it presupposes, but although these can be apprehended, noted, perceived, felt, they are pervasive qualities that like the aroma of coffee or the laziness of a summer's day resist reduction to bits of identifiable behavior.

I shall argue that the target of assessment should be the dominant mode of teaching in which the teacher is operating. In scanning this target both the explicit actions and tacit operations are examined; the explicit ones for procedural correctness; the implicit ones for their appropriateness to a given mode of teaching.

With this in mind, I shall examine the three modes of teaching to determine the aspects that might be evaluated and the criteria for doing so. It may become clear in the course of the discussion that I attribute the difficulties of the evaluation of teaching in large part to 1) the tendency to judge one teaching mode by criteria suitable to other modes, and 2) the absence of guild consensus on what constitutes procedural correctness in each of these modes.

DIDACTICS

By didactics I mean a type of teaching that tries to convey

---

*Elsewhere I have elaborated on the four uses of knowledge and the way in which Michael Polanyi's notion of tacit knowing can be used to explain the interpretive use of knowledge. Cf. "On Knowing With." *Proceedings of the 26th Annual Meeting of the Philosophy of Education Society*, 1970, pp. 89-104.

skills, knowledge, and attitudes that can be formulated explicitly with respect to outcomes and means. Matters of fact, principles, rules, slogans, manual and symbolic skills – even standard solutions to standard problems – can be classed under the rubric of didactics. The criterion for teaching-learning in didactics is the learner's ability to use the input in the solution of some non-practiced problem. Most of the standard school subjects at every level of tuition lend themselves to the didactical mode; in fact most formal instruction still consists of didactics. Standard achievement tests are the obvious measures of didactics.

One might characterize the more or less standard moves in didactics as follows:

1. Setting the task – capturing attention, making the assignment, explaining what is to be done, how it is to be done, demonstrating the criteria for a correct performance, relating task to previous learnings.
2. Inducing the pupil to attempt a trial performance of the task.
3. Correcting the trial performance.
4. Confronting the pupil with a test performance.
5. Fixing the correct performance for future recall or recognition.*

These moves with some reaching and stretching can be found in heuristics and philetics also, but not so explicitly and systematically as in didactics. More often than not, heuristics and philetics are used to conceal the didactical tactics from the pupil (pseudo-heuristics and philetics). The strength of didactics lies in its orderliness and the sense of security it gives the pupil and the teacher. It enables them to know where they are at any moment and what is expected of them. The weakness of didactics is its remoteness from the immediate concerns of the pupil. Not only is the content of standard school subjects highly abstract, but the

---

*How to carry out each of these moves constitutes pedagogy or the methodology of teaching, with which this paper is not directly concerned. Clearly, how to make each of these moves permits a wide range of variation.

very orderliness of didactics is out of kilter with the adventitiousness of ordinary life. Accordingly, for all who do not enjoy acquiring knowledge for its own sake, didactics is synonymous with the drudgery that has made school the legendary burden of the young. The history of pedagogy is replete with schemes that promise to make didactics pleasant and easy. The persistent search for such schemes is melancholy evidence that a successful one has yet to be found.*

If the five steps (or their equivalents) and their appropriate subdivisions are taken as a reasonable classification of didactics, then the evaluator can observe each one of them separately and grade the teacher on each. These gradings can be used to diagnose *loci* of difficulty, and conceivably one could devise a teacher-training component to match each step. In evaluating the execution of these moves one can note 1) the procedural correctness of each move and 2) the success of the procedures in terms of student satisfaction, student achievement, and such other criteria of success as the evaluator may wish to apply.

We now have the following possibilities:

1. The results are satisfactory or successful and the procedures correct.
2. The results are satisfactory, but the procedures are incorrect.
3. The results are unsatisfactory, but the procedures are correct.

---

*I am aware that this pessimism is scorned by those who believe that the drudgery of schooling is the mark of a teacher's ineptitude, and that with passionate interest pupils can learn anything however abstract, e.g., the batting averages of a hundred baseball players. But there is an inverse relation between the intensity of interest and the diversity of objects on which it is bestowed. There are many subjects of study, many facts, many ideas, many categories and sequences. To expect the teacher to generate passionate interest in all of them in all pupils and at all times is to place a great burden of guilt on teachers because so unrealistic an expectation is a guarantee of frustration.

4. The results are unsatisfactory, but the procedures are incorrect.

There is no problem about grading the first and fourth possibilities. What about the second one? The consumer and some of the accountabilists might rate this kind of teaching as eminently satisfactory. A profession, however, would not, for it wants to be assured that satisfactory results are not being achieved by chance or luck or simple imitation; indeed, the established professions are not satisfied even with correctness of procedures, because these can be imitated by an impostor. They want evidence for believing the practitioner knows *why* certain procedures are correct. Such understanding they infer from the study certified by the guild as necessary for entry into the guild.

The third possibility also is one that consumers and professionals will judge differently. Judging the quality of the teacher or teaching by the achievements of the learner is so plausible that it should arouse the suspicion of all who know that nothing in education is amenable to a simple solution. I shall not rehearse the inventory of variables that prevent the assertion of any simple and direct causal nexus between teacher behavior and pupil behavior. Even the most crass behavior-modification construal of teaching cannot dismiss environmental and genetic factors that might affect susceptibility to conditioning. About the best one can hope for is significant correlation between collective school inputs on the one hand and collective results on the other.

For this reason procedural correctness is the criterion to which the education profession – if there is anything one might so identify – should be committed. If, *mirabile dictu*, the five steps – or some other set – could be analyzed so that there would be substantial consensus on correct procedures, one might hope for a high degree of reliability in judgment on individual teacher competence. Furthermore, such consensus would provide the teacher with a means of defense against the charge of malpractice or incompetence brought by dissatisfied clients. Just as it is not self-contradictory for a physician to plead that his procedures were correct albeit the patient died, or for the lawyer to hold that he had conducted a brilliant courtroom defense although his

client is in prison, so must it be possible in principle for the teacher to claim procedural correctness, despite consumer dissatisfaction. The justification for all such claims, of course, (and it cannot be invoked too often) is the presence of variables that affect the results of professional procedures, but are not within the control of the practitioner. However, without a broad and strong consensus on procedural correctness, reasonable care, etc., it is virtually impossible to identify experts whose testimony would stand up in court.

The judgment on procedural correctness can be made not only on the five steps of didactics proper, but also on the tacit components of didactics, insofar as professional consensus on criteria can be achieved. In some components, e.g., classroom management, consensus on procedures might be possible; in such tacit components as class atmosphere the consensus might be no more than that "expert" judgments on global or gestalt qualities are possible. This is the sort of consensus that sustains the guild of wine experts.

HEURISTICS

By heuristics I mean a type of teaching that promotes learning by discovery or by problem solving. It is supposed to be appropriate for teaching the young how to think scientifically or civically or just logically. Process rather than product is the outcome. The Socratic dialogue is the classic example of heuristic teaching, and Socrates characterized his teaching as mid-wifery or bringing to birth ideas that presumably had grown within the mind of the pupil. Many of the curriculum developments of the 60s stressed that not content but modes of inquiry were the proper aims of education. Countless institutes were conducted to teach teachers *how to do* heuristics.

It is not at all obvious as to what the steps are in a heuristic lesson, although for one type of such teaching there is a very famous paradigm. I refer, of course, to John Dewey's analytical description of the Complete Act of Thought. In this process one can identify 1) the predicament to be gotton out of, 2) formulating the predicament into a problem, 3) entertaining an

hypothesis about diagnosing the problem, 4) deducing from the diagnosis probable remedies, and 5) applying the probable remedy to see if it really remedies.

Although these steps are rarely taken in strict sequence, they can be identified by the evaluator and presumably assessed for their logical correctness and pedagogical ingenuity. The project curriculum utilized a somewhat imprecise version of the CAT, and the steps for doing a project are quite familiar and could also be used to identify phases or steps in heuristic teaching. Nevertheless, room must be left for the creative variations that may, and one hopes will, emerge in open-ended problem solving. Hypotheses, for example, cannot be prespecified nor can they be translated into experiments for verification. Even science cannot dispense with the imaginative leap.

Unfortunately, not all types of heuristic teaching can be divided into steps, for some of them may not have a logical structure. An appreciation lesson in the arts, for example, might involve an exploration of the object for its aesthetic properties and for the depth of meanings conveyed by the images exhibited by the work of art. The teacher may guide the student's perception, hint at possibilities, assess responses, amend them or ignore them; he is guided by his perception of the flow of the discussion, not by a game plan. Or a discussion of social problems may not be ready for the logical problem-solving approach. The goal may be to search out the context for the problem rather than to decide upon a solution.

Obviously the classroom atmosphere appropriate for heuristics is different from that of didactics. The order is supplied by the problem being solved or discussed. The teacher's role is also distinctive as are the relations of the pupils to each other.* Such lessons cannot be programmed, with objectives specified, means indicated, and outcomes predicted, for the teacher may not know

*Not all heuristics are pleasant. Socrates irked some of his partners in dialogue and not all pupils can take the threats to the ego that self-examination engenders.

ahead of time what must or might happen. The process is valued for its own sake and practicing the process is the only product. What then are the criteria? The extent of the pupil's participation might be one; the sensitivity of the teacher, the inventiveness, the ability to keep the process moving in some sort of progression are possible characteristics, but, by and large, the teaching act will have to be judged as a whole by somebody who is already familiar with the quality or flavor of such patterns of inquiry.

In heuristic teaching the process is in the foreground or at the focus of attention of both pupil and teacher and didactics recede into the background. Yet it is there tacitly, for inquiries do have something to do with knowledge. Problem solving, for example, depends for its quality on the kind of generalizations pupils can exploit to form and test hypotheses. If one relies on the generalizations the groups have acquired through the informal participation in the culture, the quality of the discoveries made by the groups is likely to remain modest. Subject matter and skills learned formally through didactics tacitly structure the moves in problem solving and other modes of learning by discovery. Occasionally, in doing heuristics the teacher finds that he or she has to make didactics focal and put the heuristics into the periphery until a certain skill is developed or item or knowledge acquired. It is impossible to keep both the didactics and the heuristic moves in focus at the same time.

Even more important is the tacit operation of the teacher's command of a wide range of resources. One needs a high command of physics, mathematics, or literature to teach the modes of inquiry characteristic of them by heuristic methods. This is so because the teacher must be able to move with the questions, comments, proposals of the pupils and yet not lose sight of correctness and relevance. Heuristic teaching, properly done, is a far cry from spontaneous brain storming and the airing of opinions about whatever is suggested by the flow of the conversation.

This may also be the appropriate place to observe that trying

to achieve the outcomes of didactics through heuristics is wasteful, for heuristics is not orderly or systematic with respect to the content of a field of study or the development of a skill. Similarly, the hope that good didactics will automatically promote inquiry skills is usually vain.

In diagnosing teaching inadequacy one might look for clues that the teacher is confusing didactics with heuristics and does not know what to expect from each. The confusion may also plant in the teacher strong guilt feelings because when doing one she believes that she is neglecting the other. This guilt is exacerbated when, for example, a good didactician is employed in a school where the prevailing ideology favors heuristics or philetics. The situation becomes traumatic when a school system swears allegiance to heuristics and philetics but uses standardized achievement tests to measure teachers' competence. What one is to do with administrators, parents, and newspaper editors who confuse these modes of teaching regularly, I cannot imagine, until and unless teachers themselves stop confusing them.

PHILETICS

This term was coined from the Greek root *philos* or love. It may be used to cover all the modes of teaching, the outcome of which is an improved "relationship" between teacher and pupil, pupil and other pupils, the pupil and himself. As in other institutions, so in the school, relationships and roles are regularized and formalized by custom. The teacher, according to the tradition, was to be respected by the pupil; the teacher was to love the pupil, not as a parent or lover, but as a surrogate of the parents, the community, the culture itself. This is not necessarily a warm affective or even highly personal relationship. The problem of "relating" became acute during the 60s when the schools tried to implement the Civil Rights and the Great Society legislation. It was charged by the new humanists, social activists, and counter culturists that the teachers in the American public schools were recruited from the WASP layer of society or from those who had adopted its values with regard to work, sex, economics, and politics. These teachers, it was said, oppressed the children of the poor and of certain ethnic minorities in the inner

city.

In the jargon of the day, WASP teachers could not or would not "relate" to these children, their parents, their values, and to the environment that manifested discrimination and social injustice. Consequently, they neither expected these children to learn nor did they try to teach them. Furthermore, the argument contended, the curriculum of the public school was too bookish, too academic, and too irrelevant to the cultures of the minority groups for whom educational equality was being sought.

This is not the place to argue the merits of these arguments. They were taken seriously enough in some educational quarters to give philetic teaching the highest priority in schooling. In plain language, the teacher was to be judged by how readily he or she could empathize with the attitudes and predicaments of the children of the inner-city school. In even plainer language, it was demanded that teachers who could not "relate" to these pupils should be replaced by those who could.

Philetics in the ghetto school, therefore, is a kind of social and psychological therapy undertaken in the classroom. The teacher tries to diagnose the personality needs of the individual pupil and to meet them through serving as a parental figure, sibling figure, pal figure – as a surrogate for a person who did not play a proper role in the development of the pupil. It is hoped that by this therapy the motivational patterns of the pupils will be tuned to the demands of schooling. Central to this therapy is the demand that teachers accept their pupils and their cultures as given without passing adverse value judgments upon them. Part of the evaluation of the philetic teacher involves an ideological test, i.e., a commitment to or a high toleration of cultural pluralism.

Good philetics have always been important for successful didactics or heuristics, but always operating tacitly to provide the psychic lubricants for the learning process. To this day, being liked by pupils counteracts many deficiences in other aspects of teaching, for it is taken for granted that this will help the learning process.

Today, however, philetic teaching is urged as a focal activity with the didactics and heuristics waiting in the wings for a chance to move into focus. What, then, are the criteria for philetic teaching? Here I shall surrender to the protocols of the clinical psychologists and the guidance counselors who, presumably, have formulated approaches and steps that could be used for evaluation.

This answer is not intended to evade the issue of criteria, but it may serve to underline the fact that simply "liking" children may not be sufficient. It takes more than ordinary friendliness and interest to convince the child who is black or wretchedly poor that he is not being patronized. It takes more than well-intentioned intuition to interpret the motives and causes of suspicions and aggressiveness toward the "friendly" teacher and the well-meaning school principal. The swelling body of literature on the ambivalence of philetic relationships in the schools of the inner city with their involutions of discrimination, reverse discrimination, concealed discrimination, masked hostility, and endless rationalization render the simple prescription, "Love your pupils and accept them as human beings," simpleminded indeed. Moreover, the fact that philetics may have to be evaluated by psychiatrists, clinical psychologists, and social workers indicates the discrepancy in kind between philetics and the other two modes of teaching.

CONCLUSION

Should the evaluation of a teacher include ratings on all three modes of teaching? I suppose the evaluation of the master teacher should, but the paucity of master teachers may be accounted for by the fact that there is no high correlation among the three sets of outcomes, methods, and criteria. On most of us, fortune has smiled if we are really good at one of these modes and good enough in the other two to permit us to practice our favorite mode successfully. Consider the following examples:

For nearly 30 years class after class of high school pupils suffered under the didactics of a French teacher whose philetics were deplorable and heuristics nonexistent. She was ungainly in

body and spirit and under her ministrations French literature became as much drudgery as drill on irregular verbs. Her diagnosis of all learning difficulties was, "You are lazy and don't study enough." But in those 30 years no student who passed her courses failed the College Boards in French. On the criteria of procedural correctness in didactics she would have scored very well and by implicitly and often explicitly teaching for the College Boards she insured her success on the criterion of pupil achievement as well.

Socrates left didactics to the Sophists and spent his time teaching young men to examine their souls and the nature of virtue by cogent reasoning.* He was the heuristician par excellence, and yet this form of teaching relied heavily on the love of the young man for the excellence of the teacher and the latter's love of the pupil's youth and potentiality. It was an erotic relationship, and the Greeks made no bones about it. We tend to want the philetics without the erotics. Socrates has been rated high as a teacher but not in terms of consumer satisfaction.

I recall the day many years ago when a fifth-grade teacher told a quite indifferent scholar that she would read a little essay of his to the class. Although the teacher never knew it, the boy from that day on lived under the illusion that he had a literary career ahead of him. I would not know how to evaluate this teacher or her teaching.

In addition to the characteristics of the three types of teaching, there are global qualities of the teaching act that are discernible to the expert evaluator but not easily analyzable into procedures, performances, verbal or otherwise. I refer here to such characteristics as flexibility, quick adaptation, confidence, control, surefootedness, and the like. Teaching as a total act shares these characteristics with other human actions. Somehow they operate tacitly to guide the total enterprise, yet are not

---

*The demonstration of the slave boy learning the Pythagorean theorem by questioning in the *Meno* was not typical of the Socratic dialogue.

specific enough to isolate and judge separately. They tend not to be causes of good teaching but rather the results of success in teaching.

I have argued that we should distinguish these three modes of teaching because they have distinctive outcomes, procedures, styles. Each, therefore, has distinctive criteria and although they may combine in any given teacher-pupil encounter, at any moment only one occupies the focus; the others operate tacitly. It is this focal, explicit mode that is the proper target of evaluation.

Credible evaluation of teaching requires guild consensus on criteria, and the distinction between didactics, heuristics, and philetics may make it easier to achieve enough *consensus cognoscenti* to persuade the public and the teacher that the evaluations are not wholly subjective and idiosyncratic.

It is not impossible to formulate observable procedures for each of these modes, but there will always be knowledge, attitudes, and ideals of the teacher that will function tacitly and may or may not be observable. So in the end we cannot dispense with the judgment of the expert; the layman armed with the most minute of inventories cannot make a reliable judgment; the expert does not have to rely on such inventories. The efforts to produce expert-proof evaluators of teaching is doomed to the same fate as befell the efforts of the 60s to produce teacher-proof curricula.

I do not expect that this simpleminded analysis of teaching to clear up the mystery as to why millions of dollars and thousands of hours of research on teaching have failed to turn up indubitable criteria for good teaching, but of this I am certain: nobody – the teacher, the public, the pupil – has been served well by the expectation that a classroom teacher is being trained to perform equally well in all three modes of teaching. This expectation, based on the assumption that the three modes are highly correlated, is responsible for the odd and often contradictory sets of traits that research finds characterize "good teachers." It is also largely responsible for the perennial comedy

or tragedy of misunderstanding that ensues when a parent wants the school to teach his son arithmetic and is told that it is being done by teaching him how to think scientifically and by discovering his real identity. Above all, it is responsible for the dissonance between the rhetoric of education, which casts the teacher in the role of the charismatic heuristician and the reality, which consigns most teachers most of the time to the indispensable chores of didactics.

TEACHER EVALUATION*

Christine McGuire

Purposes of Teacher Evaluation

The goal of any evaluation is to estimate the merit, the worth, *the value* of something or somebody. It is thus more than a mere play on words to observe that the nature of the value system, implicit or explicit, from which any educational evaluation proceeds, becomes a matter of central importance in a process that necessarily eventuates in a judgment about the quality of a teacher, a method, a program, an institution, or a system. It is therefore of utmost importance to consider whose values and what values are to be applied in this judgmental process, and for what purposes a judgment is to be rendered.

In an effort to allay the anxiety which the threat of such a judgment inevitably evokes, particularly when it is made either by an outsider who may not share the teacher's value system or by a supervisor whose judgments cannot readily be appealed, some of our colleagues have argued that the value issue can be entirely avoided by the simple device of assigning to the evaluator a circumscribed technical role as collector of data and reserving to others who are often unspecified the prerogative of selecting the standards and criteria, that is, the *values*, to be applied to the evidence. It is a major thesis of this paper, however, that the value problem cannot be so easily resolved despite 1) the general consensus that for scientific purposes the *act of measuring* must be separated, at least temporally, from the *act of judging* and 2) the general recgonition that there must be competing values in any pluralistic society.

---

*Portions of this paper previously appeared in: McGuire, Christine H. "The Evaluation of Teachers and Teaching Effectiveness," in *Educational Strategies for the Health Profession*. Public Health Papers No. 61. Geneva: World Health Organization, 1974, and are reproduced here with the permission of the World Health Organization, Division of Manpower Development.

This stress on values and purposes is designed to focus attention on three philosophic issues that would appear to be of central importance in developing any model for the evaluation of teaching: 1) the nature, source, and role of values in the process, 2) the determinants of the parameters or relationships that merit investigation, and 3) the character and generalizability of knowledge that can be expected to emerge from the process. It is a second major thesis of this paper that many of the differences among currently popular models for the evaluation of teaching are attributable to divergent responses to such philosophic, rather than to "merely" technical, issues.

Finally, in addition to the philosophic and technical issues that may lead to the creation of alternate models, there is the critical matter of the functions to be served by the particular model under discussion. Ordinarily a model is developed and the evaluation process is undertaken either 1) for administrative purposes – hiring, assigning, firing, and promoting teachers, or 2) for purposes of assisting teachers in what should be their never ending quest for self-improvement or finally 3) for research purposes – to determine the relationship between teacher strategies and learner outcomes. But it should be obvious that if *research on teaching* is undertaken primarily for the purpose of increasing our understanding of relationships, *evaluation* of teaching is undertaken primarily for the purpose of making decisions. Indeed, Cronback (1964) defines evaluation "broadly as the collection and use of information to make decisions about an educational program," and suggests that, since there are many types of decisions to be made, no one set of principles of evaluation is universally appropriate.

In addition to the decisions related to individual or program improvement, there are certain administrative judgments that would seem to require models for data collection other than those implied above. With respect to these judgments, the specific question to be decided is generally of the form: "How well does X (a teacher, a student, a program, a book, a method) perform relative to specified criteria?" Or the question may be of the form: "Which one (of $X^1 \ldots . X^n$) performs best with respect to these criteria?" Or finally, it may take the form: "Is the value

received worth the cost entailed in obtaining a specified level and type of performance?" Of considerable theoretical significance is the fact that all three questions just enumerated necessarily take the form of what Gage (1963) calls a "criteria-of-effectiveness" paradigm. He observes that once the idea of effectiveness enters in, the question of a criterion is raised, and once this criterion (i.e., the dependent variable) is identified the research task is clear: "to measure this criterion, to measure potential correlates, and to determine the actual correlations between the criterion and its potential correlates." He goes on to add, "research by this paradigm has been abundant; hundreds of studies, yielding thousands of correlation coefficients, have been made. In the large, these studies have yielded disappointing results: correlations that are non-significant, inconsistent from one study to the next and usually lacking in psychological and educational meaning."

Certainly the lucid summary and insightful interpretations by Hatch and Bennet (1960) support Gate's contention that the "criteria-of-effectiveness" paradigm has failed to yield substantial results. In part this has been blamed on the inadequacies of the model, and remedy has been sought via a more sophisticated model embodying a variety of refinements. One such model reflects reconsideration of the nature of the criterion variables and entails their ordering along a temporal continuum ranging from proximate to ultimate outcomes. Predictably, the classification of potential correlates into significantly different types of categories, for example, teacher characteristics, student characteristics, teacher behavior, school environment, and so forth, and their appropriately differentiated treatment has constituted a second important approach to the refinement of the model. Among the more promising developments in this connection have been the various technical and conceptual innovations in interaction analysis that have facilitated the investigation of teaching and learning in terms of group processes. This approach has certainly encouraged careful documentation of the actual transactions that occur in learning activities and has enabled the investigator to formulate reasonable hypotheses to account for some of the inconsistent results to which Gage alludes.

Finally, there is no doubt that greatly improved research design and increasingly sophisticated statistical methodology have characterized the more recent contributions to educational research and evaluation. However, it is hard to escape the conclusions 1) that conceptual finesse in the choice of models and of statistical technique has, for the most part, been reserved for the investigation of trivia – perhaps because trivial problems usually involve many fewer variables; 2) that some of the most important questions continue to be attacked piecemeal, unsystematically, and by inexcusably crude and outmoded methodologies; and 3) that both types of studies are more often than not vitiated by the simply atrocious instrumentation commonly used in the measurement of criterion variables. Progress in instrumentation for measuring educational outcomes seems to have lagged more than any other aspect of evaluation technology; thirty-five years after the completion of the Eight-Year Study (Smith and Tyler, 1942), it is hard to point to a single major innovation in the theory or technique of instrument development, at least in the measurement of cognitive behavior.

The directions which must be taken to make maximum contributions to valid educational judgments are implicit in the above criticisms: 1) the development of a defensible theoretical construct in terms of which research and evaluation can proceed; 2) the development of a systematic schema for collecting relevant data on the important issues; and 3) the construction of criterion measures that are relevant and valid.

THE THEORETICAL FRAMEWORK

Snow (1973) observes that "the paucity of theory and research on teaching may reflect an absence of metaphor-making," and advocates a paradigm, called "The Teacher As . . ." device for promoting theoretical thinking about teaching. In this absolutely delightful essay Snow elaborates a metaphor of "the teacher as Bayesian sheepdog," in which he notes the role of the teacher in "herding" students along a path from their existing location to some presumed desired location, rounding up strays and bringing the whole "flock" together first on one side of the optimal route and then on the other. "The key features of this

metaphor are: a two-dimensional space in which a direction vector represents the teacher's chosen path and a flock of points represents pupils; a measure of average flock direction which the teacher estimates periodically; the difference between average flock direction and the teacher's direction vector, which the teacher tries to minimize; and a measure of flock density which the teacher tries to maximize. Using the pedagogical analogs of circling movements, feints, and charges, along with iterative sampling and estimation of location and direction, the teacher moves his or her class through a field, physical or psychological." (Snow 1973 p. 7)

Snow asserts that "by creating an analogy between some aspect of teaching and other known things, concepts or roles, it is possible to elaborate ideas about teaching by playing out the analogy's implications." This is what I propose to do in the balance of this paper, but because of my familiarity with medical education, I shall develop the metaphor of "The Teacher as Physician" or "The Teacher as Therapist." Fundamentally, in my view, the goals of education are not unlike the goals of therapy, i.e., to facilitate maximum functioning of the organism. Hence, the essential elements of a methodology for evaluating teachers and teaching corresponds to those employed in evaluating any other therapeutic agent, and the pitfalls of research in this area are parallel to those in any field of clinical research.

AN OVERVIEW OF CURRENT APPROACHES

In looking at the purposes of teacher evaluation (administration, self-improvement, and research) and in enumerating the necessary developments (systematic theory, systematic data collection, and systematic instrumentation), you may have noted that I had a certain affinity for trinities, hopefully, not unholy ones. Similarly, in looking at the approaches to the evaluation of teachers and teaching which are currently most widely employed and that appear to hold the greatest promise, I again suggest a trinity, this one distinguished in terms of the primary object toward which assessment efforts are directed: *perceptions*, *process* and *product*. To spin out our medical metaphor, it is interesting to note that each of these approaches has its analog in

clinical research and, like the latter, its characteristic data type and source. (Figure 1) The methods of collecting these data, their interpretations and limitations are outlined below.

Figure 1: Approaches to Evaluation of Teachers and Teaching

Object	*Data Source*	*Clinical Analogue*
Assessment of perceptions	Reports of student reactions	Evaluation of patient's subjective feelings
Assessment of process	Observations of student-teacher interaction Expert opinion	Evaluation of therapeutic rationale on pathophysiologic principles
Assessment of product, outcome or effect	Tests of student performance	Evaluation of therapeutic efficacy

Assessment of Learner Perceptions

Though the oldest, this approach is still widely used and very commonly *mis*-used. It is designed to answer the question: "How did the participants feel about the quality of the teacher, the teaching, and the program?" Formal or informal student and teacher reactions and/or ratings constitute the data for answering the question. Typically these data are collected by interview and/or questionnaire that may range from highly structured to completely open-ended inquiries. These may be couched in a form that invites the student to comment directly on his reactions, or in a form that induces him to respond to a series of indirect probes in a manner that reveals reactions about which he would be less able or less willing to be candid.

Assessment of Process

In contrast with the assessment of perceptions which deliberately seeks *opinions about the merits* of various elements in the teaching-learning situation, process assessment is designed to determine *what actually occurred* in the encounter between teacher and learner. The basic question which it attempts to answer is: "What was the nature and quality of the setting and of the communication (interaction) between student and teacher that took place in that setting?"

The information needed to answer that question consists in simple descriptive data obtained from systematic, impartial, skilled observation of the teacher and the learner in the instructional situation. Since, however, there is an unmanageably large number of attributes of the learning situation (ranging from the temperature of the room to the "heat" of the student-teacher exchange) on which observations could be made, it is necessary to select those characteristics of the instructional process which are likely to be most important in facilitating or inhibiting achievement of the desired outcomes. This decision is, at least implicitly, based on a philosophic position and a theory of learning. Further, though few adherents of this approach would admit it, the observations that are made are, in fact, designed to answer the question: "To what extent does the teacher create an environment in which the actual conditions for learning are congruent with those that *my* theory of learning assert as optimal?"

Assessment of Product

Whether a recommended set of conditions for learning is, in fact, optimal can be determined only by examining their consequences. It is to this issue that the assessment of educational product or outcome is addressed. All product assessment is designed to answer the question: "To what extent is the educational (therapeutic) agent effective in accomplishing educational (therapeutic) goals?" If it is true that teaching is undertaken mainly for the purpose of producing student learning, it follows that teacher effectiveness can ultimately be assayed only in terms of the extent to which this intended outcome is

achieved, and teacher efficiency can be judged only after the costs are documented and charged against the changes in student knowledge, judgment, skills, habits, and attitudes which are realized.

Two types of data are therefore indispensable in answering the question posed in this approach: that derived from comprehensive measures of relevant aspects of student performance, and that obtained from careful cost accounting of teacher and student time expended, other resources consumed, and undesired side effects induced by the educational program.

In looking at educational outcomes and the cost of achieving them, we may depart temporarily from the medical metaphor and look at an economic one in which the multiple outcomes to be considered include those with negative social value. Similarly, costs may be calculated in terms not only of the obvious direct costs or producing a given outcome, but also in terms of the alternate social values sacrificed by selecting one outcome or set of outcomes rather than another.

However, in applying the input-output analogy, we should note that the economic model from which it is derived imposes two important restrictions in its utilization: first, strictly speaking, the only sensible interpretation of cost is in terms of alternate products sacrificed, i.e., the "gas versus butter" argument that the cost of the former is the amount of the latter that must be sacrificed. In educational terms the analogy suggests that the costs of a new program should be calculated in terms of the amount of teacher leisure, student comfort, and other traditional values that must be sacrificed in instituting it. Second, to the economist, as to the physicist, input necessarily equals output; energy is not lost, it is merely converted; the problem resides in determining the ratio between the socially useful output and the total output obtained in the conversion process. In economic theory this problem is handled by applying a common scale of values (price) to all forms of output including those with a negative social value (e.g., smoke); but such accounting is not so easy in education where there is no generally acceptable scale for measuring the value of different kinds of output and where much

of it is even more evanescent than smoke.

THE METHODOLOGY OF CURRENT APPROACHES

Assessment of Perceptions

While the eloquence of Plato's panegyrics to Socrates has rarely been matched, his readiness to heap praise or blame on the teacher continues to be emulated. More interestingly, the teacher's willingness to invite such reaction has also continued unabated through the ages, despite the development of more reliable and valid techniques of assessing teacher effectiveness. Indeed, such testimonials constitute the sole data available about the effectiveness of most programs of continuing education in the United States and are rapidly becoming of increasing significance at both the graduate and undergraduate levels here. In their haste to develop an effective system of rewards to encourage professionalism in teaching, growing numbers of institutions of higher education are beginning to make faculty promotions and salary increases contingent, in part, on teaching effectiveness *as measured exclusively by student ratings of their instructors* – a procedure no more defensible than the determination of the therapeutic effectiveness of a drug on the basis of patient satisfaction. A society which almost weekly sees a "new cancer cure" and has repeatedly observed the rise and fall of quack therapies and medical fads that derive their popularity almost exclusively on the basis of patient "satisfaction," should know better than to rate teaching effectiveness exclusively on the basis of the victim's "Happiness index."

However, this is not to deny the relevance of student reaction data *provided they are appropriately collected and utilized*. Such basic data on student satisfaction are currently being obtained in a variety of ways: some institutions employ a standard *rating form* in which the student records his evaluation of an instructor or a course, ranking each in relation to others he has known; other institutions use a more detailed *questionnaire* designed to elicit the student's candid reaction to the several elements (e.g., a particular lecture, the required text, tutorial hours, etc.) that constitute the instructional situation. With respect to each of

these elements the student may be asked to indicate what he liked or disliked and the intensity of that feeling, what he found valuable or useless, what experiences he would prefer to have expanded or reduced, and the like.

Finally, both the rating form and the questionnaire are often combined with a third type of student reaction form in which the student is asked questions about specific feelings which actions of the instructor may have evoked in him. In general, such questions (illustrated in Figure 2) are selected on the basis of contemporary learning theory, and represent those aspects of the student-teacher relationships that are thought to influence the efficiency and effectiveness of learning.

Figure 2: Excerpts of a Participant Reaction Form

Did you feel free to challenge the instructor when you thought something he said was incorrect? Yes _____ No _____

Did you feel free to ask the instructor a question about something you did not understand? Yes _____ No _____

How often were you stimulated to think about additional applications of concepts and principles the instructor was discussing?

Practically always _____ Occasionally _____

Rarely _____ Never _____

Did you find that the instructor was sufficiently flexible in modifying his plans to suit your needs? Yes _____ No _____

Data collected from the ratings, questionnaires, and other reaction forms are summarized to show general tendencies in student evaluations. When possible, the reactions are correlated with intellectual and personality attributes of the student to

enhance their usefulness to faculty in the latter's efforts to individualize instruction.

Process Assessment

Broadly speaking, two quite different rationales and their associated methodologies characterize most forms of process assessment currently in vogue in the United States: one – the analysis of the conditions for learning – samples the totality of elements that characterize the whole of the learning situation; the second – interaction analysis – focuses only on those elements (primarily verbal elements) that characterize teacher behavior and/or student-teacher interaction.

Analysis of Conditions of Learning. Adherents of this approach begin by identifying as concretely as possible the outcomes sought in a particular instructional situation – a course, a lecture, a seminar, a laboratory, ward rounds, and the like. For example, the general outcome of "developing skill in the physical examination of the patient" could, if specified by a detailed definition of the components of that skill, be used as the starting point. Once the desired outcome has been determined, the next step in this type of assessment requires that those principles which apply most directly to the specified learning goal be identified from the vast literature on learning theory and research on learning. For the illustration given above the principles stated in Figure 3 would be among those isolated for consideration. Using such principles of learning as a frame of reference, the evaluator then proceeds to construct a highly structured questionnaire, interview, or observation form that will yield objective, descriptive data about the extent to which the "ideal" conditions for learning are met in the actual situation. An excerpt from such a form, developed for the specific illustration cited above is shown in Figure 4. Note that with minor changes in wording the questions listed there can be developed as a checklist for use by an observer, as an interview schedule for use by a trained interviewer, or as a structured questionnaire for direct distribution to students. In whatever mode they are employed the two essential elements are that each question be derived from a principle of learning that is relevant to the educational outcomes

Figure 3: Selected Principles of Learning Relevant to Learning of a Skill

Learning tends to be more efficient when the objectives are clearly understood by the learner.

The learning of a skill is facilitated by frequent, brief practice periods spaced over an extended time period.

Prompt, specific feedback that enables the learner to identify errors in his performance facilitates the learning of a skill.

Prompt, specific feedback that reinforces correct elements of a performance facilitates the learning of a skill.

Figure 4: Excerpt of a Form Designed to Assess the Conditions for Learning a Specific Skill

Were the objectives of the course (the instructional unit, the instructional exercise) explained to the students?

What is the evidence that these objectives were actually understood by the students?

How much opportunity did students have to practice *each* maneuver involved in the physical examination of the patient? How frequently were these practice periods scheduled? When (in relation to the relevant didactic material) were they scheduled? How long was each practice period? Were variable opportunities provided students in accord with individual differences in proficiency?

How often did the instructor actually observe the student examining a patient?

(continued on next page)

Figure 4 continued

Did the instructor provide any feedback to the student on the accuracy of his finding? If yes, when was the feedback provided? In what form was it provided?

Did the instructor provide any feedback to the student regarding his technical skill in eliciting these findings? If yes, when was this feedback provided? In what form was it provided?

---

sought and that each question be formulated in a manner that will yield objective, descriptive information, not opinions or preferences.

Finally, it should be noted that since communication of some kind is inherent in any instructional situation, any assessment of the conditions for learning will also include at least a few generally applicable questions that yield data about the basic characteristics of the communication. Evidence regarding this aspect of the conditions for learning would include such questions as, for example: "Could the speaker be heard by all parties to the communication?" "Did he use terms that were understandable to his listeners?" "When introducing a new term did he define it?" "Did he illustrate abstract concepts with concrete examples?" "Conversely, did he relate specific bits of information and data to a general frame of reference?" "What variety of media did he use to supplement the spoken word?" Responses to questions such as these yield the necessary data for determining whether the most elementary requisites of the communication process have been observed and, thus whether any communication could have occurred.

Interaction Analysis. In contrast with the approach described above, the various forms of interaction analysis focus exclusively on the nature of the communication process. This is accomplished by means of a detailed, systematic analysis of the behavior (predominantly verbal) of teachers and/or students. Originally developed for use in a formal classroom situation, most

forms of interaction analysis are also appropriate for use in other instructional settings. Some adherents of this approach (e.g., D. G. Ryans, B. O. Smith) focus attention primarily on teacher behavior; others (e.g., Bales, Flanders, and Amidon) use the method to examine the quantity and quality of the interaction between teachers and student.

Irrespective of the precise emphasis the general methodology is similar for most variants of this approach. Specifically, observers are trained to record at specified intervals (for example, every 3 seconds), on a form (such as that shown in Figure 5), the nature of the verbal behavior then occurring. These recordings are made for pre-defined periods (for example, 10 to 15 minutes) and the periods are selected to sample different points – for example, the beginning, middle, and end – during an instructional session or during several sessions extending over a school term. The recordings are then summarized and analyzed in terms of such variables as percent of teacher talk vs. student talk, percent of teacher talk in each behavioral category, percent of student talk in each category, characteristic sequences of teacher-student behavior, and the like.

Certain proponents of this approach (e.g., Flanders and Amidon) define the variables to be observed in strictly behavioral terms (as in Figure 5) in a manner that requires the observer to make only minimal interpretation. Others (e.g., Bales), in an attempt to get at more dynamic elements in the interaction and to identify potentially facilitative and inhibitory roles played by each participant, include such terms as "clarifies," "summarizes," "obstructs," "expresses hostility," and "challenges" in the list of behaviors which the observer is to record. Inclusion of terms such as these requires more interpretation of the observed phenomena, introduces greater subjectivity into the observations, and imposes the necessity of more extended training of observers, in order to achieve acceptable levels of inter-observer reliability. However, irrespective of the specific categories of teacher and/or student behavior selected for recording, this approach is explicitly limited to observation of the process *per se*; no attempt is made to relate that process to objectives or to outcomes, either directly or indirectly through general learning theory.

Figure 5: Excerpt of a Form for Recording Student-Teacher Interaction

Instructions: Using the symbols shown below make a notation every three (3) seconds for ten (10) consecutive minutes to indicate the nature of the behavior occurring at each 3 second interval.

Legend:

Teacher behavior:
1 = Asks a question
2 = Gives instructions
3 = Provides information
4 = Reflects student comment

Student behavior:
5 = Asks a question
6 = Answers a question
7 = Provides information
8 = Etc.

Other:
9 = Silence

Date _____ Hour _____
Class _____ Instructor _____
Observer _____ First Recording – Time _____

_____ _____

_____

_____

Now this is indeed troublesome. One could go so far as to say that the assumptions underlying some of the observational forms are frankly terrifying. The anthology, *Mirrors of Behavior* (Simon and Boyes 1967, Simon and Boyes 1970) contains 92 observational *systems* and those who have written most knowledgeably on this subject note that an additional 48 can readily be added. (Rosenshine and Furst 1973) As set up, many of these observational systems seem to imply that the voice of the student is to be treasured and the voice of the teacher abhorred. Though often treated as an axiom of interaction analysis, there is, as yet, no real documentation of this view. Indeed, McNeil and Popham observe that "effective teaching cannot be proved by the presence or absence of *any* instructional variable. (McNeil and Popham 1973) They add: "Specification of teaching skills to be displayed following training has raised an even bigger issue: The validity of the skills required. There is danger that new teachers are learning practices which are not only irrelevant, but (may even be) harmful to pupil progress. In some instances teachers are learning contradictory skills. For example, the Far West Regional Laboratory uses a minicourse to train teachers to repeat answers *less often*; but the Northwest Regional Laboratory has a training program in Flanders interaction analysis which has as one measure of the preferred 'indirect teaching' more teacher repetition of pupil answers." (McNeil and Popham 1971)

Product Assessment

Determination of Objectives. In contrast with process assessment, product assessment in its pure form is concerned exclusively with objectives (product sought) and outcomes (product realized). Product assessment, therefore, requires that the goals of the educational program be clearly specified, that they be defined in terms of behavioral changes that are to be brought about in the learner, that test situations be designed to sample these behaviors, and that these test situations be administered to the learner to determine whether he is able to perform in the desired manner and at the prescribed level of competence.

However, studies with which I am familiar monotonously reveal that the major performance factor most often measured in

the evaluation of students, teachers, and programs is the ability to recall isolated fragments of information. Such techniques of assessment would be appropriate if it could be shown that there is a high correlation between the ability to reproduce a large quantity of factual information on demand, and the ability to employ this information in problem situations. Unfortunately, such is not the case. Beyond the minimum of fundamental information prerequisite to problem solving in a particular area there is only a very tenuous relation between the ability to recall isolated facts and concepts and the ability to perform more complex intellectual tasks. In our studies of medical students at the University of Illinois College of Medicine, the correlation between scores on tests of recall and on tests of problem solving in the same content areas ranged between .18 and .41. (McGuire 1963)

These data suggest the importance of including in any student or teacher evaluation, tasks that *sample the full range of educational objectives* the program is designed to achieve, and which provide *as direct a measure as possible* of all the desired types of competence. These considerations have led inexorably to the conclusion that in order to assess the really relevant educational outcomes, it is usually necessary to design test situations that *simulate* reality as closely as possible and that require the student to exhibit the intellectual processes, the psychomotor skills, and the affective behavior that would be demanded in a real situation. At the Center for Educational Development of the University of Illinois College of Medicine we are now exploring many of these new approaches and find that a simulation technique analogous to that utilized in management games and military exercises appears to be among the most promising. In accord with these findings, we are currently developing a library of written, oral, and computer simulations of clinical problems in patient management that require sequential analysis and decision. These problems are designed to measure aspects of behavior defined by a criterion group as essential components of clinical competence. They are tightly programmed simulations designed to be suitable for either individual or group administration and amenable to either clerical or computer scoring and analysis.

The techniques to be utilized in product assessment should include the full range from conventional tests of information, through sophisticated simulations, to long-term, systematic observations of the learner in varied professional settings. These techniques do not differ in kind from those regularly employed to certify student competence. (Charvat, McGuire, and Parsons 1966) Only their scheduling and the method of summarizing their results need be modified when performance tests are administered to students for purposes of teacher, rather than student, evaluation.* Since in the former case the emphasis is on the *change* which is brought about in the behavior of the learner, it is necessary to administer the same or comparable tests both at the beginning and again at the end of the instructional period under scrutiny. Further, if long-term effects are included in the objectives of the program, a complete evaluation requires follow-up data regarding subsequent performance of the learners.

Just as the timing of any assessment is determined by the purpose for which it is made, so also must the method of summarizing the data be dictated by the purpose for which they are collected. In certifying student competence it is adequate to summarize the data in a manner which indicates merely whether the student does or does not meet all of the required performance standards. For program evaluation it is, of course, important to know how many students do or do not meet each of these standards. Indeed, if large groups persistently fail to meet any of them, this is a clear sign that something is wrong somewhere: in the teaching, the students, the test, or the standards. However, for purposes of evaluating teaching such a summary is insufficient.

---

*However, in a most thoughtful article on "Edumetric Versus Psychometric Tests," Ron Carver (*American Psychologist* July, 1974) asserts that using tests contructed in accord with usual pschometric criteria we are almost doomed to find no pre and post treatment differences. He carefully points out the pitfalls of applying such inappropriate criteria either in mastery testing or in efforts to measure educational change for purposes of program evaluation.

To serve this purpose it is necessary to summarize the performance data in a manner that helps to answer the following four questions: 1) How much *gain* was made by the total and by relevant sub-groups of students? 2) What are the characteristics of the students who appear to gain most (or least) from the current instructional program? 3) What *specific areas** are of greatest and of least gain? 4) What *specific* misconceptions (or other forms of inadequacy) are still relatively common among students at the end of the instructional program?

Cost Accounting. No product evaluation is complete in the absence of cost data, particularly since the evidence is mounting that, given adequate time, personal instruction, and self-instructional resources, most anyone of normal intelligence and emotional stability can master most any element among the objectives of higher education. The issue therefore becomes an economic one of determining whom a country can afford to educate and at what level of mastery, i.e., at what cost in resources. Given the enormous attention currently devoted in all countries to the economics of health care delivery, it is therefore astonishing how little consideration is being given in any country to the real economic issues of health professional education. Most of the data that do exist are deficient on two counts: first, they are, for the most part, retrospective; second, the categories by which data are collected are defined in such gross terms that it is usually impossible to allocate costs along genuinely functional lines.

What is urgently needed in order to evaluate teaching and teachers in cost-benefit terms is a detailed recording of the amounts of all resources that go into any instructional effort. If we are to avoid the notorious unreliability of retrospective data, this implies the necessity of undertaking diary studies of representative samples of students and faculty to determine how much of their time is actually being devoted to some kind of instructional activity, the exact nature of that activity, and the

---

*The term, "specific areas," is intended to include such diverse learning outcomes as concepts, principles, technical skills, elements of professional demeanor, and the like.

way in which the quantity and distribution of such activities change with different instructional methodologies. Second, it suggests the importance of costing out other elements in the instructional situation (such as, for example, the additional laboratory and hospital costs where patients are used as "teaching materials"). At present the only usable data tend to be those concerned with the direct money costs attributable to the construction and maintenance of specialized facilities (e.g., lecture halls, student laboratories) and materials (educational films, slides, self-instructional programs). In the absence of hard data to the contrary, it is difficult to escape the suspicion that, because their costs are hidden, we continue to cling to traditional methods that are not only ineffeetual but may also be inexcusably uneconomic since these methods rely so heavily on the most costly resource of all – professional time.

THE USES AND LIMITATIONS OF CURRENT APPROACHES

Since, if properly interpreted, some data are usually better than no data, it could be argued that the primary limitation of all current approaches lies in their inadequate usage. However, each of the approaches outlined above also has certain additional characteristic limitations which can lead to serious error as discussed below.

Assessment of Perceptions – The Fallacy that Truth is in the Eyes of the Beholder

Clearly, when perceptions are used as the primary data for evaluating teachers or teaching, there is always the danger that student reactions will be interpreted as if they were statements of facts, rather than highly subjective, individual, and variable opinions. Thus, when a student says: "The program was very valuable; I learned a great deal," there is a natural tendency to believe this to be so. Unfortunately, there is NO evidence to support the view that enjoyment, satisfaction, hard work, or any other reported "positive" reaction is necessarily correlated with learning. On the other hand, contrary to the beliefs of our Puritan forebears, there is also no evidence to support the view that something distasteful (either medical or educational) is

necessarily "good" for the recipient. Nonetheless, *provided* the data are properly interpreted as simple statements of consumer satisfactions and dissatisfactions, the approach to the evaluation of teaching through the assessment of student perceptions can be extremely valuable for two purposes: 1) to assist in the diagnosis of some of the causes of inadequate achievement and 2) to contribute to the identification and creation of conditions that will enhance motivational and other affective responses that facilitate learning.

Assessment of Process – The Fallacy of Laudable Pus

Fortunately, both students and patients are a hardy lot. Were it not so it is difficult to imagine how either could have survived the noxious interventions to which generations of both have been subjected. Consider, for example, the plight of sixteenth century patients whose physicians were severely criticized when, under their care, a wound did not develop the requsite "laudable pus" during the healing process. The censure was forthcoming irrespective of the outcome. This demand for "laudable pus" epitomizes the pitfalls of a preoccupation with process assessment independent of product assessment. For, like therapeutics, education, too, is a victim of process fads.

In an age that suffers from the information overload belched forth by highspeed computers, tireless printing presses, and armies of inexpensive reproducing machines, we find instruction disconcertingly reminiscent of the droning of medieval monks earnestly dictating to their student scribes. In an honest revulsion against this anachronism, the adherents of process assessment, though resolutely denying any imposition of preferences, sometimes seem to be operating from an implicit value system that labels the voice of the teacher as necessarily inhibitory, and that categorizes nondirective teacher behavior* as productive and

*The juxtaposition of the concepts "non-directive" and "teacher" may be puzzling to the reader; it is to this author who is simply following one of the current U.S. fashions in educational jargon.

dismisses all other teacher behavior with the perjorative – "authoritarian."

However, while educational science has been successful in demonstrating that the teacher, as such, is a must too costly resource, and an inexcusably inefficient channel to employ as a purveyor of information or a dictator of concepts more accurately recorded in readily available literature, it has unfortunately not been so successful in defining the role of the teacher positively. Repeated experimental studies of the relative effectiveness of varied instructional methodologies – lecture, discussion, programmed instruction – have produced conflicting evidence and inconclusive results. "For example, a review by Anderson (1959) of teacher-centered vs. learner-centered classroom comparisons showed 8 studies favoring the first mode, 11 studies favoring the second, and 13 studies showing no difference. Among comparisons of televised with live classroom teaching, 83 studies favored television, 55 favored conventional teaching, and 255 found no significant difference, according to a review by Schramm (1962). In comparisons of lecture vs. discussion in college reviewed by Dubin and Taveggia (1968), the tally was 45 studies on one side, 43 on the other. From a review by Baker and Crist (1971) of teacher expectancy studies, one sees that 12 studies showed no effect on student behavior and 7 showed some kind of effect. A final example comes from the Office of Economic Opportunity's study of performance contracting vs. conventional teaching. The study included some 25,000 students in grades 1, 2, 3, 7, 8, and 9 from 18 school districts around the United States. Instruction was conducted by six different companies using fairly similar teaching methods in both reading and mathematics. There are, in effect, (after losing four comparisons through missing data) 106 replications for each of two subject-matters, totalling 212. One method of analysis showed 28 significant differences favoring the programmed performance contracting treatments, 60 favoring conventional teaching, and 124 nonsignificant differences (Ray 1972). Another method of analysis yielded similar results. There were also, incidentally, scattered ATI findings in these data.

"The studies reviewed undoubtedly differ from one another in undetermined ways and contain many shortcomings; a box score

tally leaves much to be desired as a summary of findings. But one cannot fail to be impressed by the picture of contradiction that emerges. Apparently, replication by itself does not produce generalization." (Snow 1974, p. 2677)

It may have been naive to expect any definitive outcome from experiments that on reflection appear to represent a simplistic search for the educational panacea – an all-purpose, miracle method, a pursuit no less futile than the quest for that medical chimera – the all-purpose, miracle drug. However, in rejecting the role of lecturer, and in accepting the view that mature students are responsible for their own learning, some teachers have gone through what can only be described as an "identity crisis." In this anxious state many have abdicated all responsibility for defining objectives, planning the educational program, creating appropriately stimulating conditions and adequately varied resources, setting standards of student performance, and gathering the requisite data for assuring that students meet those standards. Regrettably, there has been little inclination on the part of adherents of the newer and more sophisticated forms of process assessment to solve this identity crisis by utilizing their potentially powerful tools in systematic studies of the relation between teacher behavior and student achievement.*

Until such time as the basic connections between process and product are established, the utility of process assessment must remain limited to its application as a research, a training, and a diagnostic tool. As a research tool, process analysis opens up the opportunity of determining the relative effectiveness of alternative elements in teaching strategies; as a training tool it enables the teacher to hold a mirror to his performance and to determine the extent to which his actual conduct was a true reflection of his intended strategy; as a diagnostic tool, process assessment, when used in conjunction with product assessment, yields data which

---

*In one entertaining exception, Daniel Solomon, utilizing a sophisticated factor analytic design, found the one set of teacher behaviors positively correlated with student learning could best be described by the factor name – "flamboyance."

may be of value in suggesting alternative hypotheses to explain actual outcomes that fall short of intended outcomes. At this stage of educational development, process assessment cannot be used in direct evaluation of teachers and teaching effectiveness; it can be used only in association with other methods to assist teachers in improving their performance.

Product Assessment – The Fallacy of Confusing Concurrence with Causality

Though product assessment furnishes the only data which are directly relevant in the ultimate evaluation of teaching effectiveness, it also presents the evaluator with the greatest difficulties in fundamental study design. The reasons are very simple: First, all of the factors which normally confound any research on human subjects are present in all studies of educational outcome; second, the study population is never randomly selected and its small size and lack of comparability to other populations undergoing alternative educational treatments make generalization hazardous even when conventional statistical safeguards have been applied; third, the teaching that is being evaluated usually constitutes such a small proportion of the total educational influences operating on the student subjects that it becomes almost impossible to establish plausible causal linkages; finally, the usual controlled experiment (even less, the double blind) design is rarely feasible and almost never suitable. Indeed, Robert Glaser asserts that the tyranny of the controlled experiment has probably done more to retard educational progress than any other single influence. He bases this conclusion on the observation that the only issues which can be studied by this method are so trivial they are hardly worth the effort, that the findings are available only at a point in time when they are useless for decision making, and that the requirements of the design are such that the conditions to which the findings apply can rarely be duplicated.

In addition to these inherent limitations, the approach imposes one additional hardship on the conscientious evaluator: the necessity of measuring *all* elements of product that are relevant to a particular instructional program. Assessment of a very limited range of objectives (e.g., increase in factual knowledge) continues

to be one of the most common failings of studies utilizing product assessment. This might not be too serious were it not for the fact that even the most prestigious institutions, who surely have the resources for better decision-making processes, often base educational policy on faulty, partial assessments frequently irrelevant to their major goals. A recent case in point is the report of a leading medical school which, because it fell a few ranks on a national examination designed primarily to survey the students' recall of information, abolished its experimental program and reverted to a highly traditional program. The decision may or may not have been a wise one; but it is clear that the evidence on which it was based was inadequate for the purpose.

Given the severe limitations of product assessment outlined above, it may appear that this method is of little practical utility. However, the facts are quite otherwise: *It is the one indispensable approach to the evaluation of teaching*. Indeed, McNeil and Popham (1973, p. 227) argue that "Every decision that has to do with the position of teacher and the person who is to perform the role of teacher should take into account evidence showing an *individual's competency to produce changes in learners*." (italics not those of the author) This seeming paradox is readily explained: Though it may be impossible to employ product assessment to establish that a given teaching strategy is responsible for success, such assessment is the only method by which failure to achieve the goals of teaching can be documented and the warning clearly sounded that "something is wrong." Other approaches can then be employed in an effort to identify and remedy the deficiency. The following example illustrates this point.

A CASE STUDY IN THE EVALUATION OF TEACHING*

Some years ago the author was asked to design an evaluation of an intensive weekend institute on cardiac auscultation to be

---

*A full report of the study is to be found in McGuire, C. "Auscultatory Skill: Gain and Retention," *Journal of Medical Education* Vol. 39, No. 2, Feb., 1964, pp. 120-131.

conducted at their request for the staff of a local hospital. The goals of the program were clear: to increase the accuracy of the physician's interpretation of auscultatory findings; to increase his awareness of the importance of a complete examination of the heart and systematic recording of his findings; and to improve his judgment in distinguishing innocent murmurs from potentially serious ones that justified requesting a cardiac consultation.

The evaluation design was comprehensive. It included provisions for systematic collection of student reactions, for observation of the entire program by trained personnel, and for product assessment by means of two independent performance measures to be applied at three points in time. The two measures consisted of 1) an objective test in which physicians were to identify 15 unknown heart sounds recorded on high fidelity tapes and reproduced through individual stethophones, and 2) an objective chart audit of the examination of the heart as recorded for patients hospitalized with a cardiac diagnosis. The test was administered just prior to, immediately after, and six months following the course; the chart audit was done on a series of consecutive hospital admissions for the same periods.

The findings were as follows: Students were uniformly euphoric about the excellence of the program and unanimous in reporting its value in increasing their knowledge and understanding. The accounts of the observers were equally laudatory in describing what could only be regarded as a superb instructional program. Indeed, they set aside their impartiality long enough to admit that they were tremendously impressed by the quality of the instruction. Regrettably, however, the performance measures did not support the enthusiasm of participants and observers: Though there was a statistically significant gain on the test administered immediately following instruction, this gain had disappeared by the time of the six months follow-up testing. The results of the chart audit were even more discouraging: No important differences in the examination of patients could be detected for admissions in the three time periods.

These findings led sponsors of the course to re-examine the assumptions on which it was based and to redesign it for offering

in the form of a series of short instructional sessions, spaced over a period of several weeks. At last account, this format appeared to be associated with somewhat more saluatory results.

The situation illustrated by this case is rare only with respect to the completeness of the evaluation design; the finding of dissonance between the assessment of outcomes and that of process and perceptions is not unusual and demonstrates once again the necessity of obtaining a wide variety of data as a basis for making unavoidable decisions about the effectiveness of teaching.

SUMMARY

Three approaches to the evaluation of teaching have been elaborated. The methodology of each has been presented, together with a discussion of its utility and its characteristic limitations. In light of these limitations a combination of all three approaches is recommended as indispensable.

REFERENCES

Carver, R. "Edumetric Versus Psychometric Tests," *American Psychologist* July, 1974.

Charvat, J., McGuire, C. and Parsons, V. *A Review of the Nature and Uses of Examinations in Medical Education* Public Health Paper No. 36, Geneva: World Health Organization, 1966.

Cronback, L.J. "Evaluation for Course Improvement," In *New Curricula* R.W. Heath (Ed) New York: Harper and Row, 1964, pp. 231-2.

Gage, N.L. "Paradigms for Research on Teaching," *Handbook of Research on Teaching* N.L. Gage (Ed) Chicago: Rand McNally and Company, 1963, p. 118.

Hatch, W.R. and Bennet, A. "Effectiveness in Teaching," *New Dimensions in Higher Education* Washington, D.C.: U.S. Office of Education, 1960.

McGuire, C.H. "The Evaluation of Teachers and Teaching Effectiveness," WHO/EDUC Series 73/165, Geneva, 1973.

McGuire, C.H. Auscultatory Skill: Gain and Retention, *Journal of Medical Education* Vol. 39, No. 2, Feb., 1964, pp. 120-131.

McGuire, C.H. "Research on the Process Approach to Analysis of Medical Examinations," *1963 Yearbook* National Council on Measurement in Education, 1963.

McNeil, J. and Popham, J. "Assessment of Teacher Competence" In *The Second Handbook of Research on Teaching* Travers, R.M.W. (Ed) Washington, D.C.: American Educational Research Association, 1973, p. 233.

McNeil, J. and Popham, J. from Rosenshine, et al. "Research in Teacher Performance Criteria." In *Research in Teacher Education: A Symposium* B.O. Smith (Ed) Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

Rosenshine, B. and Furst, N. "The Use of Direct Observation to Study Teaching." *The Second Handbook of Research on Teaching* Travers, R.M.W. (Ed) Washington, D.C.: American Educational Research Association, 1973.

Simon, A. and Boyes, E.G. (Eds) *Mirrors for Behavior: An Anthology of Classroom Observation Instruments* Vol. 1-6, 1967 and Vol. 7-14 and Summary, 1970, Philadelphia: Research for Better Schools.

Snow, R.E. "Representative and Quasi-Representative Designs for Research on Teaching" *Review of Educational Research* Vol. 44, No. 3, Summer, 1974, p. 277.

Snow, R.E. "Theory and Construction for Research on Teaching," *The Second Handbook for Research on Teaching* R.M.W. Travers, (Ed) Washington, American Educational Research Association, 1973.

EVALUATING TEACHING

by

Irvin J. Lehman

When I first accepted the NSPER Co-directors' invitation to prepare this manuscript, I underestimated the complexity of the task. Some of the relevant research findings are contradictory. There is disagreement on those traits or characteristics of the good teacher. There is disagreement about the validity of evaluative methods. Student ratings are a case in point. Evaluating teaching is not as simple as it appears on the surface. Like a mosquito in a nudist colony – I don't know where to begin. But an assignment was accepted – an effort must be made. Whether or not I have been successful in my attempt to clarify the issues inherent in evaluating teaching, will be a matter for you to decide. I will put theory into practice, so to speak, by having the reader evaluate my teaching effectiveness: Did I state my objectives? Did I have an enjoyable style? Did I communicate effectively? Did I encourage the formation of stimulating questions? These are just some of the myriad questions that an evaluator of teaching might consider.

To start with I want to make explicit an assumption to which I subscribe. Teachers and teaching can be evaluated. Perhaps not as precisely or elegantly as some other aspects of education, or, as carefully controlled as a chemical experiment. But, teachers and teaching can and are being evaluated. Further, those evaluations, if carefully conceived and controlled, can be used to improve instruction and enhance learning. I further submit that systematic improvement of teaching and learning is virtually impossible in the absence of evaluating teachers and teaching.

I do not approach this task unaware of the criticisms of and abuses in the evaluation of teaching, particularly those emanating from the use of student ratings. At the same time, however, I am unable to accept arguments such as, "infringement on academic freedom," "de-emphasis of research," etc. If we stop and think for a moment, the teacher performs his or her professional role in

the absence of other professionals. How often do we see only one surgeon or physician in the operating room? One lawyer in the courtroom? One welder on the assembly line? All other professionals perform in an open arena. So, is it not a legitimate demand of decision makers, be they students, administrators, parents, or taxpayers, to expect, nay, demand that teachers also operate in an open atmosphere and be subjected to some form of accountability? We can no longer consider the classroom as a sacred entity and the teacher as one whose effectiveness should not be measured.

There are six major acts in my scenario:

Act I: What do we mean by evaluation?
Act II: What do we mean by teaching?
Act III: Why evaluate or what is the value of evaluation?
Act IV: What should be evaluated?
Act V: Who should do the evaluation?
Act VI: How should the evaluation be performed?

The play unfolds in the classroom, the home, the administrator's office, the school board meeting room, and on the floor of state legislatures. In each of these settings the scene has commonalities and variability hopefully dealt with in the composite of the six acts.

Beyond these six acts the paper summarizes research designed to determine characteristics of effective teaching and closes with a discussion of rating scales in evaluating teaching.

WHAT DO WE MEAN BY EVALUATION?

Evaluation, according to the dictionary, "is to determine the worth of; to appraise." Cronback (1963) defines evaluation as the collection and use of information to make decisions about an educational program. These decisions may be related to course improvement, about the individual pupil, or administrative actions, about how good a school system is, or about teachers and new programs. Stufflebeam (1973) says evaluation "is the process of delineating, obtaining, and providing useful information for

judging decision alternatives." According to Stufflebeam, evaluation involves making value judgments about the worth of several competing alternatives, information needed to choose one alternative over the other(s). Stake (1967) portrays evaluation as "describing and judging an educational program." Scriven (1967) defines evaluation as "gathering and combining performance data with weighted sets of goal scales." Other definitions have been provided by Provus (1969), Tyler (1942), and Alkin (1969), to name a few. Although the definitions vary, they all appear to share a common purpose – providing information to the decision maker to assist him in making educational decisions.

Although evaluation and educational research have much in common, they nevertheless are different and this should be kept in mind. The major purpose of educational research is to add to our knowledge of the educational process. The major purpose of evaluation is to provide a basis (information) for making decisions about alternatives. It is true that research sometimes is the medium whereby the information needed to make the decision is obtained, but this should not be interpreted as saying that a research study is an evaluation study. Table 1 illustrates the differences between the "ideal" research and evaluation study.

Simply put, research is that activity which is concerned with the study of the relationships existing among variables to permit generalizations. Its ultimate goal is theory building, and it may or may not have immediate application. Evaluation, on the other hand, has (or should have) immediate application and is concerned with determining the worth or value of a product, program, or procedure.

WHAT DO WE MEAN BY TEACHING?

Teaching is commonly defined as the process, technique, or procedure used by a person to induce behavioral changes in another person. Normally speaking, the person inducing the change is referred to as the teacher while the person in whom the change is to be induced is called the pupil. However, evaluating teaching requires further clarification. Are we concerned with the process, the product, or both? If we are interested in the process,

Table 1: Comparison Between "Ideal" Research and Evaluation Study

In an "ideal" research study	*In an "ideal" evaluation study*
1. Researcher selects the problem.	1. Problem is determined by a particular situation. The study is undertaken in response to a need to know the utility of an alternative.
2. Hypotheses are generated by either deduction or induction.	2. Precise hypotheses can't be generated. Rather, the evaluator must rely on judgment and experience to "fill in the voids" in knowledge.
3. Value judgments are limited to those implicit in the problem.	3. Value judgments are made explicit in the problem selected.
4. Can be replicated.	4. Nearly impossible to replicate because the problem is typically unique to a particular situation.
5. Problem determines the type of data to be collected.	5. Feasibility of data collection often determines the type of data collected.
6. Possible to control relevant variables.	6. Control difficult. At best, only superficial control.
7. Primarily concerned with the generalizability of data, discovery of new principies, theory building.	7. Primarily concerned with gathering information about the value of an educational activity so as to help make decisions about improvement and management.

do we mean the instructional method, e.g., lecture, discussion, lecture-lab, lecture-quiz, primarily lecture with little discussion, primarily discussion with little lecture? By process do we mean the manner in which the course is organized? The course structure? The substance or depth of the course? If we are interested in the product of teaching, what are our criteria? Are we talking about behavioral changes? If so, which ones? Are we talking about pupil performance vis-a-vis pretests, posttests, and retention tests? Are we concerned with behavioral changes in cognitive skills or in affective behavior? Should one of our criteria be future success of our graduates? Should we be concerned with cost-benefit analysis? These are just some of the myriad of questions that must be considered by those interested in studying some aspect of teaching.

A more perplexing issue, possibly, is of a philosophical nature. To speak of defining teaching by examination of its product implies the existence of clearly defined goals. Presently, our schools, colleges, and universities do not have clearly formulated goals. Those that do exist are presently in a great state of flux. For example, although most college administrators would agree that the role of higher education is not to provide the student with a "union card," much of the basis for recruiting students in the past has relied on this dogma. With students experiencing the exasperating fact of more and more unemployed college graduates, they are beginning to question the real value of a college education. Unlike a few years back, colleges cannot guarantee that once graduated you will be in a certain category and income bracket. Hence, using this as a possible criterion (success of graduates) for ascertaining the effectiveness of a teacher or institution is precarious to say the least. This therefore implies that we ask a different set of questions such as: How much is the student learning? What is the value of his education? What does his education really mean? To answer these questions we must begin to focus our attention on the *products* and the *processes* that make up that nebulous quantity which we call an education. This situation, coupled with a tightening of state and federal funds for all levels of education (and I see it worsening before improving) has resulted in a closer scrutiny of the effectiveness with which educational institutions, regardless of

their level, meet both their objectives and society's goals (in the ideal one hopes that they are one and the same). In an attempt to focus on the outcomes of education, administrators are beginning to look to their faculty in an attempt to assess their effectiveness and the faculty evaluation movement is here. But, I am getting a little ahead of myself. I'll come back to this at a later point.

WHY EVALUATE OR WHAT IS THE VALUE OF EVALUATION?

As mentioned earlier, there are mounting pressures from students, parents, taxpayers, and other decision makers to hold our education system accountable. These pressures, which will become even stronger as our economy continues to move toward recession, constitute one answer to the question, "Why evaluate?" They provide information on the quality of the job being done by those who participate and those who pay for it.

There are other, perhaps more compelling, reasons for evaluating the educational enterprise. These reasons are as follows:

1. Broadly speaking, we evaluate to ascertain the degree to which the teacher's or schools' instructional objectives have been realized.
2. Teachers should be evaluated to help identify the effective and ineffective teacher. Judiciously used, the identification of the ineffective teacher should *not* be made solely for obtaining evidence to fire that teacher. Rather, especially if the teacher has tenure, evaluation-type data should be used to identify the teacher who might be given appropriate instruction so as to help that teacher become more effective.
3. The results obtained from properly conducted evaluation studies can assist the decision maker in matters relating to promotion, granting of tenure, merit salary increases, and the like. I would be remiss, however, if I did not issue a word of caution in using evaluation results as the *sole* criterion upon which to make any type of decision. No doubt the fact that evaluation studies are used too often as the major

criterion in certain kinds of administrative decisions explains, in part at least, the negative attitude, the hostility toward, and the fear of many teachers toward evaluation of their teaching ability, regardless of the means used.

4. We evaluate to provide teachers information which hopefully will result in creating an optimum teaching-learning environment. Naturally, the type of information gathered should be related to the instructional process. Another word of caution is appropriate now. The teaching-learning process is very complex. It involves a myriad of factors, many of them interacting at any given time, some of which have yet to be isolated. As yet, research has not been able to provide us with clear-cut data that will permit us to differentiate between the effective and ineffective teacher. (See Gage, 1963, for a full discussion on research on teaching.) At the same time, there are some factors, which on the surface, might appear to be totally unrelated to the instructional process. And there are some questions asked on teacher rating scales related to certain teacher traits or characteristics that have been found to be unrelated to the instructional process.

5. Education is our single greatest expenditure at the state and local level. Since Sputnik, massive financial support has been given to the development of new science and mathematics programs such as BSCS, SMSG, and PSSC. We now have numerous Title I and Title III projects funded at the federal level to help provide pupils with the best education possible. We have television programs such as "Sesame Street" and "Electric Circus" to improve the experiential background of disadvantaged children. Are these new programs worth the time and effort devoted to their development? Those persons supporting the adoption of new courses and programs, or those considering their use should demand some type of evaluation which will demonstrate the effectiveness of the newer technique or program or product over the older one. If we, as professional educators, don't ask for it, nay, demand it, legislators, parents, and critics (and we sure have a potful in educational matters) will. I'm not afraid of their demands. Rather, I am fearful of the kinds of information they will want and how they will use (or should I say, misuse) it. The same need emanates from recent technological advances such as

teaching by TV, tape recorders, filmstrips, etc. These and growing emphasis on individualized instruction and computer-assisted instruction (CAI) create a need for information on their effectiveness.

6. Evaluation results used to judge the effectiveness of a school, district, state, regional, or even national educational system can have great influence on the institution being studied. While the results may not have much influence directly on the students, they can have a profound influence on the curriculum, the staff, and the administration.

In summary, we undertake the evaluation of teachers and teaching for a variety of reasons – to help administrators make decisions; to help teachers make decisions that will help them provide an optimal teaching-learning environment so as to maximize student learning; to help those responsible for making budgetary decisions (how much money should be allocated and where it should be distributed); to help students with their decisions. And if for no other reason, we should undertake evaluation studies related to the teaching-learning process for self-defense.

We have many critics of our educational enterprise. Everybody is an expert on teaching. Our critics claim that we are spending too much for education, especially on some of our newer programs. They contend that our children are not learning as much today as they did in the "good old days." They complain that our students are irresponsible, apathetic, and ill-informed about the world in which we live. It is conceivable that our critics might be justified with their comments for certain schools, classes, students, or programs. I contend that we have a *good* (I didn't say nor do I think that it can't be improved) educational system. However, to obtain the data needed to silence or to deal with our critics – be they students, parents, legislators, or school board members – about the quality of education, evaluative information is needed. Innovation for innovation's sake is not enough. We must prove the worth of our innovations. The Research and Policy Committee of the Committee for Economic Development (1968) aptly stated it when they said:

> Innovation in education, whether it involves the use of new curriculum materials or new educational technology, has become essential if the schools are to be genuinely effective in achieving their aims and goals. Continuing assessment of the product of schools also is necessary. This means the development of principles and techniques for critically judging the worth of whatever the schools teach and the effectiveness and efficiency of their methods of instruction.

Evaluation is a two-edged sword which can be either beneficial to the personality development of students and enhance learning or be very destructive. It can either have positive effects on various educational decision makers or have negative effects on the curriculum or school system.

We recognize that there are weaknesses inherent in any evaluation technique – even the ones that have been carefully planned and meticulously conducted. Therefore, it behooves us as teachers of those who will at some time in their lives be involved in evaluating educational efforts and products, to design our studies in such a way that they will have a positive effect on both students and teachers by providing them with information that will optimize teaching and learning. For this reason, we must be cognizant of the following:

1. Evaluation studies that focus on trivial and insignificant factors will have little, if any, effect on the teaching-learning process. It is true that we have many gaps in our knowledge of the teaching-learning process. However, we do know that there are some factors that are completely irrelevant to it.
2. Evaluation studies that are not perceived by students as providing information for instructional improvement are likely to have minimal effects. For example, if Mr. Jones' students know that he seldom or never looks at the results of a teacher rating scale, we can expect that these students will develop a negative attitude to rating scales and may even generalize their feelings to all instructors, even those that diligently use the results for instructional improvement.

3. We must dispel some of the negative notions of teachers toward the benefits to be derived from faculty evaluations. It is one thing to be against student-conducted course ratings which may be used indiscriminantly, are invalid because of a limited sampling of student opinion, or are invalid because of the criteria being measured. It is another thing to be against the utility of teacher evaluation in general.

4. Teacher evaluation information should be the personal property of the teacher being rated. Unless the teacher knows that his superiors will not see the results and possibly use the information for nefarious purposes, teachers will be hostile toward them.

WHAT SHOULD BE EVALUATED?

If one subscribes to the definition of teaching as that activity which induces change in the students' behavior, he should accept the proposition that the ultimate criteria of effective teaching are changes in students toward the instructional objectives of the school. Now comes the nitty-gritty question. What are the school's goals? Without wishing to go into a prolonged discussion about the goals of higher education, we should note that they involve *both* affective as well as cognitive changes. And, in the elementary grades, they might well include psychomotor changes. In many instances, the less obvious changes in students' beliefs, ideals, interests, values, and motivation, to name a few, are changes that colleges hope to affect. Too often, people think only in terms of cognitive changes. Has Jim learned how to solve a quadratic equation? Has Eileen learned the concept of valence? Does Allan know the parts of speech? I submit that the inculcation and accumulation of knowledge is an important aspect of one's education since it provides the basis or foundation upon which to build the higher mental processes. At the same time, I feel that developing critical thinking ability, an appreciation for the arts, and desirable attitudes are also important and, if taught, should be subjected to some sort of evaluation.

In contrast to the evaluation of the school's product, there is another answer to the question, what should be evaluated? That

answer is process – how does the teacher operate in his or her efforts to achieve the products? There are several items to be considered in the evaluation of process: role, style, techniques, and methods. There are various conceptions about the teacher's role in the classroom. Some contend that he should only be a dispenser of knowledge. Some feel that he should be a counselor. Some feel that he should be a socializer and help students develop socially acceptable conduct and attitudes. Some contend that the teacher should be a motivator, a disciplinarian, a substitute parent. There are differences in perception of the teacher's role depending upon whom is asked – students, teachers, or parents. Which one is right? Which should be evaluated? Unfortunately, there is no right or wrong answer, for it depends upon the purpose of the study and the use to which the evaluation results will be put. As mentioned earlier, evaluation of teaching can be used for course improvement, decisions about individuals, and for administrative regulation. Evaluation of teaching can be undertaken by the individual faculty member, and sometimes in cooperation with his peers to provide him with information that will assist him in building strengths and eliminating weaknesses. On the other hand, quite distinct from self-assessment, administrative evaluations are made by members of the administrative team for purposes such as retention, promotion, and tenure. Too often, however, the same type of information is gathered – which is unfortunate. Just as there is no such thing as universal validity, there is no all-purpose teacher rating scale.

As far as evaluating the teacher's role is concerned, the evidence to date suggests that no single source – teachers, parents, pupils, administrators, and other educational decision makers – should be relied on; that the perceptions of all persons involved in the educational process must be considered in defining and hence evaluating the teacher's role.

The evaluation of instruction must, by its very nature, address itself to the teacher's style, that is, the teacher's personality. A teacher's style not only has an effect upon the instructional techniques he uses in the classroom but in some instances may be related to his effectiveness.

Finally, any evaluation of teachers and teaching must consider the teacher's technique. Technique as used here is *not* synonymous with method. Rather, it refers to the strategies – reward, punishment, praise, criticism – used by the teacher to accomplish his/her instructional objectives. Research has shown that there is a significant relationship between a teacher's style and the effectiveness of his/her instruction.

To answer the question "What should be evaluated – the teacher's role, style, technique, or product?" I submit that *all* of these should be considered. If there is to be a variation in emphasis on these things, the use of the data gathered – that is, the kinds of decisions to be made – will determine the focus of the study and hence the traits to be evaluated.

WHO SHOULD DO THE EVALUATION?

If we subscribe to the definition of a good teacher as one from whom a student learns that which he should learn and define good teaching as the process used to provide maximum learning, the student who sits in a particular teacher's class for the course should be in the best position to provide *certain* kinds of information. That student should be able to react to questions about the course structure, organization, quality and quantity of assignments, the teacher's enthusiasm, the manner in which the teacher dealt with questions, the teacher's reactions to views differing from his own, the substance of the course, and the like. The student is in the best position to know whether he gained anything as a result of taking the course, whether he would recommend the course to others. The student is not only an excellent source of these types of data, he is also the most direct, the primary recipient of the instruction being performed. Thus, the student is one answer to the question, "Who should evaluate teaching?" I do not think however, that students can or should be expected to evaluate teaching if the criteria relate to publication record, administrative ability, or ability to work with his colleagues.

Evaluation can have both a positive and negative effect on the teaching-learning process. If judiciously interpreted and used, the

data can provide valuable feedback to the instructor. Based on the student's reactions to his course structure, instructional objectives, tests used, assignments made, and instructional method(s) used, the instructor hopefully will be able to obtain information that will assist him in future decisions regarding his teaching and may suggest modifications so that maximum teaching and learning will result. The results of evaluation studies, if *mis*used can also have marked effects on the teaching-learning process, but in a negative direction. Thus, a second answer to the question is "the teacher."

As mentioned earlier, there are mounting pressures from students, parents, taxpayers, and other decision makers to hold our educational system accountable, pressures which must be responded to by school administrators. These pressures will become even stronger as our economy continues in its recession. There is no denying the fact that instruction and teachers at all levels will be evaluated with more vigor although possibly for the wrong reasons. I submit that teachers, their instructional strategies, and products will be evaluated by both persons in and out of our schools and colleges. Therefore, lest we as professionals lose our grasp as the major agency to conduct evaluation studies, it behooves us to pay attention to the laments and comments of our citizenry – students, parents, taxpayers, and legislators, to name a few – be they right or wrong in their criticism; be their concerns for constructive or destructive purposes. And, it behooves them to be evaluators, to be additional answers to the question, "Who should evaluate teaching?"

Each of these concerned groups has a stake in the education system, in the quality of its performance, and in the decisions which direct it. If we ignore the need for, and if we leave evaluation to the courts, radical students, disgruntled parents, taxpayers, and legislators who are balking at education's spiraling costs, we will regret it and, like Hamlet, forever have the ghost of doom haunt us.

I stated at the outset of this paper that numerous parties *are* involved as evaluators of teaching. Besides students listed above we can quickly name teachers, administrators, parents, board

members, etc. I do not want to argue here whether all of these parties *can* or *should* engage in such evaluations. At a very minimum, *teachers should* evaluate teaching to help them improve their own instruction; *students* should evaluate teaching inasmuch as they are in the best position (sitting in the class) to know about the teacher's style, organization, grading, etc.; and *administrators* should evaluate teaching to improve school programs.

WHAT DOES RESEARCH TELL ABOUT VARIABLES USEFUL IN THE EVALUATION OF TEACHING?

Whether one is interested in program, curriculum, product, or process evaluation or whether one is interested in evaluating the teacher, per se, will be governed by the purpose of the study. If you were interested in determining the extent to which your pupils have mastered a particular unit in 11th grade physics, you would engage in product evaluation. If you were interested in ascertaining whether Method A is better than Method B in teaching a particular concept in the unit on physics, you would undertake process evaluation. The remainder of this section presents a brief review of my reading of what research says about effectiveness of teaching methods. Its purpose is the delineation of the characteristics of teaching. Thus concerns about process, product, summative, and formative evaluation are for the moment irrelevant. Rather, we only concern ourselves at this time with considering the research findings as related to various instructional techniques.

Instructor versus student-centered instruction. Research comparing these methods show mixed results. One study reported that students in the student-centered class had a significantly higher grade although the students felt that they would have done better in an instructor-centered class. The students felt that they had received greater social and emotional value from the student-centered discussion groups but that even though there was more opportunity for discussions, there was lack of direction and little attention paid to errors by the instructor (Faw 1949). Another study (Asch 1951) reported contradictory results – the students in the instructor-centered classes received significantly

higher grades, while the students in the student-centered classes felt that they would have learned less in an instructor-centered class. Generally, the method used in the presentation of the course materials has little perceptible influence on achievement (Landsman 1950; Birney and McKeachie 1942).

In respect to the affective measures, it appears that the student-centered method may be slightly more advantageous than the instructor-centered method (Ruja 1951; Gibb and Gibb 1952). Students in the group-centered classes appear to make significant gains in self-insight, in independence, and in their interpersonal attitudes. There is also much evidence that group-centered instruction produces important social and emotional changes (Johnson and Smith 1953). However, factors other than the method of instruction seem to play an important role in determining the nature of the findings, and the number of these factors is so great that many have yet to be studied (Jamison et al. 1974).

Lecture-discussion. Studies comparing the discussion method with other methods of instruction show no conclusive evidence regarding the value of one method over another on conventional achievement measures (Ruja 1951; McKeachie 1951). Students prefer the discussion method over formal lectures but would rather have a combination of the two to either method alone. Research has also shown the discussion method to be better than the lecture method when problem-solving skills are to be developed (Bloom 1953; Drake and DeLulio 1954).

Laboratory method. Again, no perceptible difference is shown between the laboratory or demonstration method and other teaching methods except insofar as manipulative skills are concerned (Kruglak 1954).

Independent study. Students in independent study programs learn at least as much as students taught by the conventional approach as judged by performance on achievement tests.

In summary, these studies as well as a thorough review of the literature by Dubin and Taveggia (1968) indicate that at the

college level there is no distinct superiority of one teaching method over another.

Class size. There is little evidence to support the commonly held assumption that large classes are ineffective and that the quality of instruction is universally related to the student-teacher ratio. Whenever any differences were found, it was in studies dealing with young children (here small classes appear to improve cognitive and affective development).

Instructional television. There is no conclusive evidence regarding the relative superiority of traditional teaching method over ITV or vice versa, although more positive results are obtained at the lower grade levels.

Programmed instruction. Again, I find no evidence to suggest the superiority of programmed instruction over conventional approaches. It *may* result in students achieving specific instructional objectives.

Computer-assisted instruction. At the elementary level, when CAI is used as a supplement to a traditional approach, it has been found that there is an improvement in the pupils' achievement, especially in the slow-learning students. At the secondary school and college level, the evidence does not suggest that traditional instructional methods are better. However, because it may result in substantial savings in student time, release the teacher for other activities, and from a cost-effective study is less expensive, we may find CAI and programmed instruction becoming more popular in the next few years.

Student-faculty opinions. Student and faculty reactions to the various instructional methods are mixed. Both teachers and students are more favorable to ITV at the elementary level than in the secondary school and college. Administrators are more likely to favor ITV than teachers. At the college level, students prefer small classes to television classes but prefer television classes to large classes. Reactions to televised courses vary according to the subject being learned. Faculty vary in their opinion regarding one approach over the other. Students are more

receptive to ITV *after* they have experienced it than before. Student approval or preference for a certain teaching method is related to the students' personality characteristics. No doubt the same would hold true for teachers.

The implication of the research reviewed above is that successful teaching cannot be assured by the adoption of some proven method. Research on instructional methods, class size, and use of various media shows no distinct advantage of o..e method or approach over another in regard to learning outcomes. One might say, "One teacher's discussion is another teacher's lecture." Any research conducted in this area *must* consider the subject matter, the instructional objectives, the personality and background of the teacher and student, the educational environment, and the materials of instruction. It is the proper balancing of all these factors that yield good results. Individual rather than general solutions must be found. Thus the need for developing means to evaluate instructional practices on an individual basis must be emphasized.

The many variations of traditional instruction (lecture, discussion, laboratory, independent study) have been subjected to extensive observational analysis within the past decade, especially in our elementary and secondary schools. More recently, because of the "budget squeeze" some of these studies have been replicated at the college level. A most striking observation garnered from these studies is that students learn effectively from all these media (be it traditional instruction, CAI, PI, or instructional TV) and that with few exceptions there is no significant difference in one medium over another or of one variation in a medium over another. It was also seen that little, if anything, would be gained insofar as pupil performance is concerned were a school to introduce some newer approach. The few exceptions to this were in the area of class size at the primary grades. Here smaller classes *might* improve the students' performance.

One might think that to determine which of two instructional methods is better requires that some type of research design be developed wherein one group of children would be taught by one

method, the other group by the other method. Then the results of the two groups would be compared and the group of children receiving the higher scores would be said to demonstrate that this method is better. Is it that simple? No! How do you know whether gain or loss in learning might not be accounted for by the novelty of the learning experience? How much improvement or loss in learning might be due to the students' and/or instructors' motivation? How easy would it be to walk into a school and request that you be able to randomly assign pupils to treatments? What about the criterion problem? What about the interaction of methods, student characteristics, teacher characteristics, and other variables? Can individual behavior be inferred from group behavior? These are just some of the factors which, as you well know, must be considered.

Evaluation of different teaching methods cannot and should not end with just a study of the differences in achievement test scores between the various instructional groups. We must also consider whether there might be behavioral changes other than cognitive skills, whether improvement had been obtained at the expense of achievement in other courses? Should we not consider transfer of training to other courses taught in the traditional manner? Should we not consider behavioral changes evident immediately after the instruction but which do not last? Should we not consider the possibility of a latent effect, that is, changes in behavior which may not manifest themselves until a period of time has elapsed? Yes we should. But do we?

Ideally, a study of the effectiveness of teachers and teaching would relate input measures, the educational process, surrounding conditions, and some set of output measure(s). In a sense, we would end up with some type of regression equation. Naturally, studies would have to exhibit control. Although it might not be too difficult to randomly select subjects for the various treatments, the major problem in such studies is the large number of potentially relevant variables that have to be controlled. And, if a researcher tried to control for these relevant variables, he might end up having no control. Those of you interested in designing experiments to compare different teaching methods should consult Jamison et al. (1974), and Schramm (1971).

EVALUATING INSTRUCTION USING RATING SCALES

Assuming that the function of evaluating teaching is to provide for an optimal teaching-learning environment, student evaluations using some form of student rating scale have been found to provide valid and reliable information. In fact, research findings, vis-a-vis evaluating teachers and teaching, indicate that the validity of student ratings is not affected by such things as sex, class size, rater's G.P.A., major or non-major status of the rater, grade expectation, academic level of the student rater, severity of grading, and the like. And, researchers have found that most of the instruments are fairly valid and reliable. Because of the problems associated with other techniques, I strongly recommend using student instructional rating scales to evaluate teachers and teaching.

Let us now consider several common methods of evaluating instruction via rating scales. Basically, there are five major types. They are normally classified as *descriptive, numerical, graphic, forced-choice, and man-to-man*.

In the *descriptive* scale, the rater is asked to check the one response which he feels to be *most* descriptive of the instructor or the method of instruction. Usually, several traits are included in a single rating scale and the rater checks the statement that is most descriptive for each trait. Following is an example of a descriptive item:

I believe my instructor

_____ presents his material very well
_____ has good presentation techniques
_____ is about average in his presentation of material
_____ presents his material poorly
_____ presents his material very poorly

This approach could be used to obtain the students' reactions to the teacher's knowledge of his subject matter, course objectives, sense of humor, attitudes toward questions and discussion, personal relationship with students, course

assignments, grading practices, and other factors.

In the *numerical* rating scale, the rater is required to indicate along some continuum that usually ranges from 0 to 10 (although the range is arbitrary and personally I like a maximum of 5) where a particular trait would fall. For example, the descriptive example given above thus could be modified:

> Directions to rater: Below is presented a variety of statements pertaining to your instructor and his method of teaching. Rate each of the statements from 0 (very poor) to 10 (excellent) with 5 being the average.
>
> _____presentation of material
> _____attitudes toward students
> _____knowledge of subject matter
> _____realization of course objectives
> _____organization of class material
> _____critical analysis of assignments
> _____fairness of grading practices used
> _____relevancy of assignments

The *graphic* rating scale is similar to the descriptive and numerical types, especially the latter since the student is required to rate a particular trait along some continuum. The difference lies in the fact that the ratings are made in a graphic form which affords the user a pictorial presentation of the traits being evaluated. Some graphic rating scales such as the Purdue Rating Scale for instruction provide for the presentation of mean scores of each instructor on a summary form which expresses the mean score in terms of percentiles. An example of a graphic rating scale formerly used at Michigan State University is shown in Figure 1.

In the three types of rating scales already considered, the rater considered only one trait at a time and either selected the most descriptive behavior of the ratee for a particular trait, assigned the ratee some numerical score, or placed the ratee along a continuum. In the *forced-choice* rating scale, the rater is

Figure 1

MICHIGAN STATE UNIVERSITY
TEACHER EVALUATION SHEET

Major _____ *Sex* ____ *Class* _____ *Grade Point Average* ____

DIRECTIONS: It is the desire of your instructor to achieve the best possible instruction in this course. To help accomplish the purpose, this evaluation sheet was devised to obtain a systematic poll of student opinion. Carefully consider each question, then record your judgment by encircling one of the letters A, B, C, D, E, for each item. A blank space has been provided at the end for adding comments you wish to make.

1. Were important objectives met?	The course is an important contribution to my college education	Contributes about as much as the average college course	This course doesn't seem worthwhile to me
2. Does instructor's presentation of subject matter enhance learning?	Presentation very meaningful and facilitates learning	Presentation not unusually good or bad, about average	Presentation often confusing: seldom helpufl
3. Is instructor's speech effective?	Instructor's speaking skill Concentrates my attention on subject	Speech sometimes invites attention on speaker rather than subject	Speech usually distracting, concentration very difficult

4. How well does the instructor work with students?	I feel welcome to seek extra help as often as needed	I feel hesitant to ask for extra help	I would avoid asking the instructor for extra help unless absolutely necessary
5. Does the instructor stimulate independent thinking?	Instructor continually inspires me to extra effort and thought beyond course requirements	In general, I do only the usual thinking involved in the assignments	I seldom do more than rote memory work and cramming
6. Do grading procedures give valid results?	Instructor's estimate of my over-all accomplishment has been quite accurate to date	Instructor's estimate of my accomplishment is of average accuracy	I feel that the instructor's estimate is quite inaccurate
7. How does this instructor rank with others you have had?	One of the best instructors I have ever had	Satisfactory or about average	One of the poorest instructors I have ever had

Comments:

(favorable)

(unfavorable)

required to choose from a series of statements, one or more statements most descriptive and one or more least descriptive of the person or instructional method being rated. This is analagous to some interest tests that you might have taken where you are required to select from among three activities the one you liked best and the one you liked least. From prior statistical analysis, quantitative values can be assigned to each of the choices. Development of such a scale is a complex task. However, reduction of rater bias, elimination of some errors of response set, and a more normal distribution of scores make this a promising approach. One must be careful, in constructing such a scale, to make certain that all foils are plausible. Following is an example of a forced-choice rating scale:

> Directions to rater: Below are a series of statements which one might use to describe your instructor and/or his teaching method. In each set of four statements, pick out the *one* that is *most typical* or descriptive of your instructor, and the *one* that is *least typical* of your instructor or his method. Place an "M" before the most typical and an "L" before the least typical description in each of the series of statements.
>
> _____ is patient with slow learners
> _____ volunteers to help students
> _____ is well prepared with his lectures
> _____ makes clear and meaningful assignments
> _____ is fair in his grading
> _____ encourages questions
> _____ encourages discussion
> _____ is tolerant of opinions differing from his (her) own

Of the four foils in each series, two, on the basis of prior item analysis, should have been selected because they distinguish between good and poor instructors. The remaining two may be attractive but they are innocuous and nondifferentiating. Scores can be assigned to each of the foils, and the instructor's score would be the sum of scores obtained for fifteen or twenty such series of four statements.

The *man-to-man* rating scale has been used extensively in our armed forces for rating officers and non-commissioned officers. It is, however, adaptable to rating instructors. The man-to-man scale consists of three to five items which are descriptions of particular persons or personality types, each of whom is well-known to the rater. The rater then uses these persons as "frames of reference" and places each of the subjects to be rated along a scale defined by the "standards" group. Only one trait can be rated at a time and hence, with a multiplicity of traits, the man-to-man scale can be most time-consuming. And, since college students or those pupils above the sixth grade seldom have the same instructors for all their classes, it is not often feasible to use this method for students to rate their instructors. It is usable in peer ratings, especially at the departmental level, although the faculty might not approve of it. Following is an example of a man-to-man scale:

> Directions to rater: Listed below are several traits regarded as desirable for a good college instructor. Following each trait are the names of five people you know personally. For each trait, choose the person who in that regard is most like the person you are rating. Place an "X" in the space beside the name of that person.
>
> Trait being rated: SCHOLARSHIP Ratee: Professor Ingle
>
> _____Prof. Doe
> _____Prof. Gephart
> _____Prof. Lehman
> _____Prof. Craig
> _____Prof. Donovan

We have already noted that faculty, regardless of their level, are not too receptive to having themselves or their instructional method rated, especially by students. This is so even though, properly conducted rating scale based evaluation studies can provide an instructor with valuable feedback information that could be used for course improvement, that students might be able to obtain more insight into their own role in learning and

accept more responsibility for it, and that administrators could use the information as *one* of many pieces of information upon which to make decisions related to hiring, promotion, granting of tenure, etc. An indirect way of approaching faculty ratings might be via *course ratings* rather than instructor ratings which relieve somewhat the negative faculty reactions to items with direct personal implications.

In developing such a scale, a very large number of statements are placed on cards which are then sorted by faculty members into groups on a continuum from highly desirable to highly undesirable. From those statements upon which considerable consistency in judgment is found, a subset representing the entire continuum is selected. Each statement is then assigned a scale value which is calculated from its position relative to other statements. Thus, a composite score can readily be obtained for any course. The technique illustrated here is called the Q-sort. Below, is an example of a general course rating scale which is part of the Purdue Rating Scale series.

Name of course _____

Directions: Check (x) every statement below that expresses your feeling about this course.

( ) 1. This course is very beneficial to the majority of students who take it
( ) 2. This course does not help students much
( ) 3. I am very enthusiastic about this course
( ) 4. This course is too vague

Another approach to teacher and instructional method ratings would be to use an open-ended or free response format. This form is *not* recommended because of the difficulty in summarizing the free responses. Its particular merit is that it forces the student to consider his own role and responsibility in the teaching-learning process. The instructor, by reading the responses, obtains reactions which can have obvious implications for how he conducts his classes. If, for example, the majority of student responses indicate a dissonance in percep-

tions between students and instructor with regard to course objectives, it is evident that the instructor has failed to communicate them to his students.

There are, as we have seen, a variety of rating scales that can be used to evaluate teachers and teaching. We have also stated that all of them are subject to some form of error, some more so than others. Also, the nature of the items might vary from one evaluator to another. These types of rating scales differ in response format used. They all typically contain items dealing with course content, course organization, and the depth or substance of the course.

CONSTRUCTING RATING SCALES

Although it is not my intent to provide you with a formal exposition of the construction of rating scales (those interested should study the chapter by Remmers in Gage, et al. 1963), I would like to point out a few of the more salient points of which you should be cognizant.

1. Identify a set of behaviors thought to be associated with effective teaching. This step in essence defines the content or the types of items to be used in the rating scale. Obviously what is relevant may vary from one evaluator to another. But generally speaking, the knowledge of the teaching-learning process, the purpose of the study, the opinions of others, and the critical incident technique are some sources that will guide the evaluator.
2. Identify *a priori* a group of teachers deemed to be effective and ineffective.
3. Have many more items on the preliminary scale than will actually be used. You will be surprised at the large number of items that "bomb out."
4. Having different groups of students rate various instructors and use cluster or discriminant analysis to identify those characteristics that are most discriminating insofar as the two groups of teachers are concerned.

In using rating scales, as with any other evaluation device,

validity and reliability are of prime importance. The various forms discussed may or may not be valid in particular circumstances, depending on their content and on the manner in which they are administered. At best, student ratings are valid indications of student opinion; they may or may not be valid as indicators of the quality of instruction. Students, contrary to general opinion, are inclined to be very charitable, to the point that little discrimination may be found among the ratings of different instructors. Aleamoni and Hexner (1973) reported that students who were informed that the results of their ratings would be used for consideration of the promotion and advancement of the instructor rated the course more favorably than students not given this information. This, of course, raises some very serious questions. Should students be given this information? Might the manner in which the rating scales were administered influence the validity of the ratings obtained? The answer to the latter question is yes. The findings strongly suggest that the prior instructions given the students are an important confounding variable that must be considered when interpreting teacher evaluation ratings.

In some institutions, rating scales are administered to the students in the instructor's absence. This practice, as far as some faculty members are concerned, "stinks." Many students might be overly liberal to avoid inflicting personal embarrassment and possible professional injury on the instructor.

Whatever the ultimate disposition of the results, it is desirable to develop an atmosphere of receptivity so that the faculty will cooperate fully in administering rating scales to students to insure anonymity of response (there are a few, fortunately very few, instructors who might seek retribution and hence invalid responses will be obtained) and to make clear that frank and honest appraisal is desired as a basis for improving instruction. The use of rating scales as a "stick" by department heads and deans can only be injurious to faculty morale. It also encourages in students a false sense of their ability and importance in passing judgment on their teachers. Rating scales too often emphasize the thespian qualities of a teacher rather than his role as an initiator of the learning process. Education is a

two-way street, and when students evaluate their teachers, they should also rate their personal contributions to the learning process.

It is my personal belief that the results of rating scales should be shared with no one unless the instructor being rated so desires it. I feel that the primary value to be gained from rating scales is information that the instructor can use to become a better teacher. Unless instructors feel confident that neither nefarious comparisons nor injudicious decisions will be made by administrators, I am fearful of the validity of student ratings. My remarks here should not be construed as meaning that administrators should not evaluate their faculty. They definitely should. However, they should employ other means to gather the needed information.

Should students be provided with the results of faculty evaluations by means of rating scales? This is indeed a difficult question to answer and no doubt, regardless of my feelings, I am most certain that I will have colleagues disagreeing with my point of view. I answer with a *qualified* yes. I make this qualification purposely. If the student ratings have been obtained using a valid instrument, I see no reason why students should not have the opportunity to select instructors who other students have rated as good teachers. The students are paying for their education (albeit a small proportion, but they are still paying something toward it) so are they not entitled to get the best instruction possible? I think so. I can assure you that it won't take too long for students to recognize whether the results of faculty ratings are invalid.

Much of the ineffectiveness of rating methods lies, as mentioned earlier, in the selection of criteria of effective instruction. Some traits are amenable to rating, some are not. Examples of the former are efficiency, scholarship, perseverance while such traits as humane, kind, unselfish are most difficult to rate (Wrightstone 1960). As indicated earlier, the abundance of research conducted to identify characteristics of the effective teacher has yielded very few useful findings.

Another problem with rating scales is the language used. An effort must be made to use words that have the same meaning to all raters. "Scholarship" for example, may mean different things to different students. The chemistry instructor who is able to "rattle off" complex formulas may be termed as a scholar whereas the art professor who has developed a new theory may be called an "egghead" by a chemistry major who has to take an art elective. The dilemma of vocabulary might be ameliorated somewhat by using the critical-incident technique. After students have rated, described, or chosen words from a given list, they might be asked to describe a specific instance of an instructor or classroom activity which illustrates this quality. The student's basis for appraisal might also be had by asking him to name his best and worst instructor and explain by giving an actual experience with that teacher, why he rated the instructor as he did. Since students react to overt behavior more than to esoteric personality and character traits, rating scales should be so worded to reflect this. Moreover, character traits are likely to be ambiguous and their proper combination for effective teaching is unknown.

A commonly voiced faculty concern and ensuing lament is that student ratings by and large are not too valid, that is, student ratings may be influenced by teacher traits such as severity of grading, amount of homework assigned, and grade received in the course. Research has shown that student ratings are little influenced by grades received in the course, by sex, by class size, and by course difficulty. Research has also shown that the instructor's popularity, showmanship, and benevolance have no bearing on the ratings students give that instructor. Research has also shown that alumni, ten years after graduation, agree substantially with on-campus students in the ratings given the *same* instructors. The most extensive studies on the relationship between the *student's classs* and *teacher's rank* showed a tendency for graduates to give higher ratings than undergraduates and for instructors to receive the lowest rating by both than undergraduates and graduates. There was, however, no consistent pattern evident for the other ranks (Clark and Keller 1954; Riley, et al. 1950; Remmers and Elliott 1949). Another study by Guthrie (1954) showed substantial

agreement between the ratings given by graduates and undergraduates as well as a high degree of correlation between student and peer ratings of the *same* instructors.

The observational method, especially that termed naturalistic observation (and rating scales can be classified as one type of observational approach), is a viable approach to studying teacher effectiveness. It does, unfortunately, as most methods do, have problems associated with it. Some of these problems are: 1) obtaining valid, relevant data about a complex behavior which is often subjective, 2) making sure that highly correlated behaviors are not measured at the same time lest a particular behavior(s) be weighted disproportionately, 3) obtaining at least two observations by at least two trained observers, 4) having an adequate sampling of behaviors for an adequate sampling of subjects (time sampling may help to insure that infrequently occurring behaviors will be observed), 5) developing an observational schedule that will contain relevant aspects of the behavior being studied, and 6) observer bias. With respect to observer bias, we must recognize that there can be rating errors or response set biases such as "halo effect," "error of leniency," "error of central tendency" (assigning an average rating to all traits for all subjects studied), and "contamination" (where the rater knows, for example, those teachers who are considered good and those considered poor before he rates them). There are of course other examples of observer or rater bias such as having a Black researcher observe subjects to see if they display prejudice towards Blacks and the effect of the observer(s) in the classroom on the behavior of those being observed and rated (Hawthorne effect).

All the research evidence to date on student ratings suggest that valid rating scales do have a place in evaluating instruction. Students are perceptive, and they become more so when they realize that their opinions are seriously regarded. Neither student nor, for that matter, any other rating scale data should be used as the final index of the quality of instruction. Some recognition of their value is suggested by the fact that nearly 25 years ago, about 40 percent of our colleges used student ratings. This figure no doubt is much higher today. A survey of

evaluation practices in 468 school systems in 1973 showed that student evaluation of teachers was reported by more school systems than any other form of client-centered evaluation.

OTHER MEASURES OF INSTRUCTIONAL EFFECTIVENESS

The lengthy consideration of rating techniques should not cause us to lose sight of the fact that there are other ways of evaluating instruction. As mentioned earlier one can appraise the results of instruction through: 1) the use of achievement test score gains,* 2) noting the extent to which accepted instructional functions are carried out (motivating the student, setting standards of performance, providing the student with some satisfaction in his progress, etc.), 3) noting the number of students selecting an area as a major or seeing the number of students who take additional courses in an area, 4) using an observational approach of interactional analysis, e.g. Flanders' and Amidon's scales, 5) noting the number of students who select Professor X as their major professor or academic advisor, 6) seeing the number of students who try to take courses (enrollment) with Professor X, 7) correlating grades with ability, and 8) comprehensive examinations. Each approach has obvious difficulties if overemphasized as the single criterion for effective instruction.

SUMMARY

I have attempted to discuss with you the various methods by which instruction and instructors can be evaluated. And, throughout, I've attempted to show the value to be accrued from sound evaluations. All the evidence gathered to date has not demonstrated the relative superiority of one instructional method over another. Indeed, it is unlikely that it ever will since there are a myriad of factors which must be considered in

---

*See Anderson, 1972; Bormuth, 1970; Glass, 1974; Stake, 1971 for a discussion on using standardized test scores to evaluate teaching.

evaluating various teaching methods or teaching devices or teachers. It would appear that the characteristics of the teacher, students, and subject matter, and the nature of the course objectives have as much or more effect on the quality of teaching and learning than does the class size or the method of teaching.

Several methods for evaluating instruction were presented and discussed: peer ratings, student ratings, course ratings, etc. Because of the elusive nature of the teaching-learning process, no completely adequate mode of evaluation exists, but there is evidence that properly constructed rating scales are valuable aids in evaluating instructional methods and instructors, especially when the emphasis is placed on the improvement of the quality of instruction.

In closing, let me say that I am not disheartened by the fact that we, as educators, have a large void in our knowledge of evaluating teachers and teaching. But we do have some information and we should continue our search for better approaches. We shouldn't become dissuaded in our search.

Educators should be decision makers! Every classroom teacher, principal, and superintendent should have a good knowledge about evaluation of instruction if only so that they can be an intelligent consumer of the various journalistic diatribes on the quality of education in 1974. We must realize that evaluation, especially accountability is here to stay and will become more potent in the future. Therefore, we need qualified personnel to assuage the vilifications uttered by some of our critics. I have discussed where we are in the "state of evaluation of instruction." I think that you will agree that teaching in 1984 ain't gonna be like teaching in 1974. The next ten years may be turbulent from the educator's point of view. But one thing is crystal-clear in my mind. We can't hide our heads in the sands like an ostrich and hope that things will get better. We can, and I am confident that we will rise to the challenge of the next decade. To do a better job will require dedication, hard work, perseverance, a strong heart, and the ability to weather the winds of adversity which will come our way, sometimes in

tornado force. I am confident that we will rise to the occasion and weather the storm. I would like to conclude with a little saying as well as a little prayer:

> All our lives long, every day and every hour, we are engaged in the process of accommodating our changed and unchanged selves to changed and unchanged surroundings; living in fact, is nothing else than this process of accommodation; when we fail in it a little we are stupid, when we fail flagrantly we are mad, when we suspend it temporarily we sleep, when we give up the attempt altogether, we die.

I submit that if we are not prepared to accommodate, we had better get out of the noble profession that we are in. Hopefully, a little prayer is in order here.

O LORD

Give us the strength to accept with serenity the things that cannot be changed,

Give us the courage to change what can and should be changed,

Give us the wisdom to distinguish one from the other.

I am confident we will overcome!

REFERENCES

Aleamoni, L.M., and Hexner, P.Z. "The effects of different sets of instruction on student course and instructor evaluation." Office of Instructional Resources, Urbana, Ill.: University of Illinois, 1973.

Alkin, M.C. "Evaluation theory development." *UCLA/CSE Evaluation Comment*, No. 2, 1969, 2-7.

Anderson, R.C. "How to construct achievement tests to assess comprehension." *Psychological Monographs*, No. 3, 1951.

Bassin, W.M. "A note on the biases in students' evaluations of instructors." *Journal of Experimental Education*, 43:1, Fall 1974, 16-17.

Birney, R., and McKeachie, W. "The teaching of psychology: A survey of research since 1942." *Psychological Bulletin*, 52, 1955, 51-68.

Bloom, B.S. "Thought processes in lectures and discussion." *Journal of General Education*, 7, 1953, 160-169.

Bormuth, J.R. *On the theory of achievement test items*. Chicago: University of Chicago Press, 1968, (mimeo).

Caro, F.E. (ed.) *Readings in evaluation research*. New York: Russell Sage Foundation, 1971, 1.

Centra, J.A. "Effectiveness of student feedback in modifying college instruction." *Journal of Educational Psychology*, 65:3, 1973, 395-401.

Clark, D.E., and Keller, R.J. "Student ratings of college teaching." In Eckert, R.E., and Keller, R.J. *A university looks at its programs*, (eds.) Minneapolis: University of Minnesota Press, 1954, 197-212.

Cronback, L.J. "Course improvement through evaluation."

Teachers College Record 64, 1963, 672-683.

Drake, R.M., and DeLulio, A.M. "Improvement of Instruction." *Review of Educational Research*, 24, 1954, 305-312.

Dubin, R., and Taveggia, T.C. *The teaching-learning paradox: A comparative analysis of college teaching methods*. Eugene, Ore.: University of Oregon Center for the Advanced Study of Educational Administration, 1968.

Dyer, H.S. "Toward objective criteria of professional accountability in the schools of New York City." *Phi Delta Kappan*, 52:4, 1970, 206-211.

Faw, V.A. "A psychotherapeutic method of teaching psychology." *American Psychologist*, 4, 1949, 194-199.

Festinger, L. *A theory of cognitive dissonance*. Evanston, Ill.: Row, Peterson, 1957.

Glass, G.V. "Design of evaluation studies." In Worthen, B.R., and Sanders, J.R., *Educational evaluation: Theory and practice*. Worthington, Ohio: Charles A. Jones Publishing Company, 1973, 225-232.

Gibb, L.M., and Gibb, J.R. "The effects of the use of participation action groups in a course in general psychology." *American Psychologist*, 7, 1952, 247.

Guba, E.G. "Evaluation and the process of change." In *Notes and working papers concerning the administration of programs*. Title III ESEA. Washington, D.C.: U.S. Senate Committee on Labor and Public Welfare Subcommittee, 1967.

Guthrie, E.R. *The evaluation of teaching: A progress report*. Seattle, Washington: University of Washington, 1954.

Jamison, D., Suppes, P., and Wells, S. "The effectiveness of alternative instructional media: A survey." *Review of Educational Research*, 44:1, 1-68.

Johnson, D.M., and Smith, H.C. "Democratic leadership in the college classroom." *Psychological Monographs, No. 11, 1953.*

Kruglak, H. "The laboratory and purposeful activity." In French, S.J. (ed.) *Accent on teaching.* New York: Harper, 1954, 197-201.

Landsman, T. "An experimental study of a student-centered learning method." Syracuse University, 1950, Unpublished doctoral dissertation.

Provus, M.M. "Evaluation of ongoing programs in the public school systems." In Tyler, R.W. (ed.). *The sixty-eighth yearbook of the National Society for the Study of Education*. Chicago: University of Chicago Press, 1969, Part II, 242-283.

Remmers, H.H. "Rating methods in research on teaching." In Gage, N.L. (ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963.

Remmers, H.H. and Elliott, D.N. "The Indiana college and university staff evaluation programs." *School and Society*, 70, 1949, 169-171.

Research and Policy Committee of the Committee for Economic Development. *Innovation in education: New directions for the American school*. New York: Committee for Economic Development, 1968, 13.

Riley, J.W., et al. *The student looks at his teacher*. New Brunswick, N.J.: Rutgers University Press, 1950.

Ruja, H. "Non-directive teaching and self-insight: A statistical addendum." *Journal of General Psychology*, 51, 1954, 331-332.

Scriven, M. "The methodology of evaluation." In Tyler, R.W. (ed.), *Perspectives of curriculum evaluation*. Chicago: Rand McNally, 1967, 39-83.

Schramm, W. *Notes on instructional cross-media comparisons*. Stanford University Institute for Communication Research, (working paper) 1971.

Smalzried, N.T., and Remmers, H.H. "A factor analysis of the Purdue rating scale for instructors." *Journal of Educational Psychology*, 34, 1943, 363-367.

Stake, R.E. "The countenance of educational evaluation." *Teachers College Record*, 68, 1967, 523-540.

Stake, R.E. "Evaluation design, instrumentation, data collection, and analysis of data." In Worthen, B.R., and Sanders, J.M. (eds.), *Educational evaluation: Theory and practice*. Worthington, Ohio: Charles A. Jones Publishing Co., 1974, 303-316.

Stake, R.E. and Denny, T. "Needed concepts and techniques for utilizing more fully the potential of evaluation." In Tyler, R.W. (ed.), *Sixty-eighth yearbook of the National Society for the Study of Education*, Chicago: University of Chicago Press, Part II, 1969, 370-390.

Stufflebeam, D.L., "Educational evaluation and decision-making." In Worthen, B.R., and Sanders, J.R. (eds.) *Educational evaluation: Theory and practice*. Worthington, Ohio: Charles A. Jones Publishing Co., 1974, 128-142.

Tyler, R.W. (ed.) *Sixty-eighth yearbook of the National Society for the Study of Education, Part II*, Chicago: University of Chicago Press, 1969.

Wrightstone, J.W. "Observational techniques." In Harris, C.W. (ed.), *Encyclopedia of Educational Research*, New York: Macmillan, 1960.

THE TEACHER APPRAISAL FOR IMPROVEMENT (TAFI) WORKSHOP

The following is a description of the Center for the Study of Evaluation's *Teacher Appraisal for Improvement* (TAFI) workshop as used by the NSPER: 74 participants.

1. Objectives – There are three objectives that the workshop is designed to enable its participants to achieve: a) that participants will be able to list the major components, advantages, and disadvantages of current methods of teacher appraisal; b) identify the techniques and procedures that can be used to improve these methods; and c) apply them in meeting their own teacher appraisal needs.
2. Participants – A minimum of 30 and a maximum of 50 participants is recommended. The participants should include administrators and school board members when appropriate, i.e., everyone who might be concerned with and/or use the information gained from a teacher appraisal system. No special training is required to participate in the workshop.
3. Duration – The workshop takes one day to run (6½ – 7½ hours), although it is sometimes preferable to allow up to two days. The longer time period permits more group discussion of the pros and cons of the various methods discussed, especially in relation to their applicability, utility, and installation in the participants' schools.
4. Materials – Each participant receives a notebook containing information about the various kinds of teacher appraisal systems discussed, along with the directions for the various team exercises. Each three member team of participants receives a team notebook containing the team exercises and answers to these exercises for each participant. At the conclusion of the workshop the participant's notebook serves as a reference tool on teacher appraisal methods.
5. Content – Eight techniques and methods that may be useful in the appraisal of teaching are presented. Brief examples of their application are presented and their strengths and weaknesses are outlined. These eight are:
   a. Unstructured Observations and Ratings – (Subjective and idiosyncratic observation of teaching by adminis-

trators, peers, or students)

b. Systematic Observations and Ratings – (Use of pre-specified procedures for observing teaching and recording descriptions of teaching behaviors: e.g., Flanders Interaction Analysis, teacher skill checklists, etc., need by administrators, peers, or students)

c. Teacher Skill Testing – (The presentation of a teaching-learning situation via a written or verbal description or an audio-or videotaped presentation and asking teachers to prescribe instructional strategies appropriate to the situation)

d. Omnibus Measures of Student Growth – (The use of standardized tests to determine amount of student growth related to general educational goals and thus to infer quality of teaching)

e. Objectives-based Measures of Student Growth – (Testing students on specific objectives concentrated on during periods ranging from several weeks to a semester)

f. Very Short Measures of Student Growth – (Testing students on the acquisition of facts, principles, or skills taught in a brief instructional period)

g. Teacher Performance Testing Systems – (Structuring a teaching-learning situation concentrating on a single objective and observing teaching and learning related to that objective)

h. Analysis of Typical Student Growth – (The use of regression analysis to identify effective teachers and subsequently the techniques used by teachers whose students perform at or above predicted achievement levels)

6. Organization and Content – The workshop contains two phases. In the first phase, the participants are grouped into three-person teams. The teams receive instruction about a particular teacher appraisal method, e.g., ratings. This instruction is provided by means of the participants reading a certain section of their notebooks and/or the workshop leader presenting a brief lecture on the topic and/or a structured discussion of the topic. At the conclusion of the instructional period, the participants complete a team exer-

cise. These exercises provide "hands on experience" in solving realistic problems associated with the exercise. Feedback is then provided as to how the exercise problem should have been solved. There is then a discussion of these answers and the potential applicability and utility of the method for the needs of the participants. This sequence of instruction-exercise-feedback-discussion is repeated for each of the following general kinds of teacher appraisal methods: ratings, teacher skill competency tests, and methods based on student performance. Within each of these general methods, several approaches are considered. For example, within the category of ratings, the workshop considers ratings (by peers, experts, students, and outsiders) in conjunction with various strategies for performing such ratings. These latter strategies range from impressionistic judgments to systematic and quantified data gathering on specific student and teacher actions.

The second phase of the workshop is quite different than the first in that there is a complete simulation of the methods considered in the first phase. This involves random assignment of participants to the roles of "teachers," "students," or "expert observers." Specially prepared packets are distributed to participants to aid them in role playing the various roles in this major simulation. A short lesson is actually taught and the results of the various appraisal methods are statistically compared, e.g., the average difference in the pre- and post-scores of the "students" are correlated with the average ratings these "students" gave their teacher. At the conclusion of this exercise, the applicability and utility of the various methods are again discussed in light of the participants' better understanding of what is really involved in their use.

At the conclusion of the simulation exercise, the participants complete and discuss a questionnaire on which they rate the method or combination of methods that they feel would be useful in their local setting.

Some of the exercises require two teams to work with one

another, and in one exercise, one participant leaves his "own" team and works with two others. This structure helps the participants to interact and share their experiences.

7. Evaluation – Three 10-item tests covering the workshops' objectives are administered as a pretest and again as a posttest. Under this procedure a given participant takes a different 10-item test for the posttest than he/she took for the pretest. This item sampling approach allows for all 30 items to be administered to each team on both the pretest and posttest, but without spending too much time in the collection of this data. A post-workshop questionnaire is also administered to all participants in order to get their opinions of the quality of the workshop.
8. Staff – The staff required for this workshop includes a discussion leader, knowledgeable in the area of teacher appraisal, and a coordinator (e.g., a secretary, graduate student, or paraprofessional) to assist in logistics during the second phase of the workshop. The responsibilities and activities of both the leader and coordinator are thoroughly detailed in a leader's manual which accompanies the workshop.

SOME COMMENTS AND OBSERVATIONS

The Teacher Appraisal for Improvement Workshop well received by the NSPER: 74 participants. It provides for a variety of learning experiences, many of which are designed to allow participants to simulate some of the activities involved in teacher appraisal. Furthermore, the eight techniques presented, although not exhaustive, are representative of the state-of-the-art in the evaluation of teaching.

Two caveats are in order, however. First, the workshop's title "Teacher Appraisal for Improvement" may be easily misinterpreted. The workshop focuses on APPRAISAL for improvement, not on the TECHNIQUES FOR THE ACTUAL IMPROVEMENT of instruction. This seems based on the assumption that before we can systematically improve instruction we need to determine the current quality of that instruction. TAFI concentrates on the methods useful in assessing teaching.

Once that assessment is accomplished a decision can be reached as to whether instructional improvement is needed. If it is, information and techniques not dealt with in the TAFI Workshop must be employed to create the needed instructional change. Second, the third objective of the workshop – that participants will be able to apply the eight methods in meeting their own teacher appraisal needs – might create some unrealistic expectations. Participation in the workshop would not transform an individual (otherwise lacking knowledge and experience with the selection and/or construction of instruments and the transformation and use of the data obtained) into a qualified evaluator. It would enable participants (e.g., teachers, administrators, and school boards) to be informed consumers of evaluation information obtained through these methods and to knowledgeably participate in the design of a school's plans and procedures for evaluating teaching.

There are three points that should be clear in the minds of TAFI Workshop participants at the end of the session. If they do not surface in the discussions structured into the workshop schedule, the leader should present them in a wrap-up session. Those points are:

1. PURPOSES FOR EVALUATING TEACHING. There are at least two purposes for evaluating teaching that will surface during the workshop: a) for helping the teacher improve his or her skill as a teacher, and b) for helping administrators to make decisions about promotion, assignment, salary, and retention. TAFI FOCUSES ON THE FORMER. It is true that some of the methods discussed in the TAFI Workshop can be used for the latter purpose. BUT, there will be differences in their use and in the overall plan if the focus is on the administrative purpose.
2. SELF-EVALUATION OR EVALUATION BY OTHERS. This point, who does the evaluation, is not directly considered in the TAFI Workshop. Given the purpose of evaluation of teaching to assist the individual to improve instructional skill, it is clear that many of the techniques dealt with in TAFI can be used in either a self-evaluation or an evaluation-by-others format. It is also clear that

some of the techniques rest heavily on one of these approaches. This self vs. others question must be faced in designing the overall evaluation plan for a specific setting.

3. NEXT STEPS. The completion of a TAFI Workshop *does not* give the participants an evaluation plan. Nor is TAFI an approach to evaluating teaching. Rather, TAFI presents an array of techniques that have been used in evaluating teaching, discusses their individual strengths and weaknesses, and suggests that some set of them might be used in a school's plan for evaluating teaching. But it does not present that plan. It only gives some of the information that should be considered as decisions are made about what should be included in such a plan. The next step is up to the people in a school and it is the creation of the evaluation plan that will serve their purposes.

This report concludes with a summary of the NSPER: 74 participants' reactions to the TAFI Workshop.

DATA SUMMARY AND ANALYSIS FOR THE TEACHER APPRAISAL FOR IMPROVEMENT WORKSHOP CONDUCTED BY THE NATIONAL SYMPOSIUM FOR PROFESSORS OF EDUCATIONAL RESEARCH

NSPER Workshop sites:

Denver, Colorado	October 21, 1974
Seattle, Washington	November 4, 1974
Atlanta, Georgia	November 18, 1974

Pretest – Posttest Results

There are three forms of the multiple-choice achievement test (Forms A, B, and C). Each test form contains ten items for a total pool of thirty items. Workshop participants receive different forms of the test for the pretest and the posttest.

The pretest – posttest data were positive for all three workshops. A summary of the results is presented below:

NSPER Workshop Site	N	T value	Significance Level
Denver, Colorado	55	2.57	p .01
Seattle, Washington	48	4.67	p .01
Atlanta, Georgia	83	1.90	p .01

Participant Questionnaire Results

The workshop materials are accompanied by a questionnaire that is designed to ascertain the participants' attitudes towards the workshop's content and their perceptions of the instruction. Responses to the questionnaire revealed some interesting information. In general, participants at the NSPER sessions indicated that they:

1. would use the information learned at the workshop in their jobs;
2. would recommend the workshop to others;
3. would be interested in participating in similar workshops

dealing with other aspects of evaluation; and

4. found active involvement in the exercises helfpul in stimulating their learning.

Workshop participants recognized the utility of the TAFI information. For example, the question, "Will you use what you have learned from the workshop in your own job?" elicited this response (based on a total N of 198 for the three NSPER sites): 67%, yes; 4%, no; 28%, not sure.

An overwhelming majority of respondents (N=196) for the three sites would recommend the workshop to others: 88% recommend, 12% not recommend. Also a substantial proportion of participants would be interested in participating in similar workshops dealing with other aspects of evaluation (N=203): 85%, yes; 5%, no, 10%, not sure.

The instructional style of the TAFI workshop was well received by NSPER participants as evidenced by the following data (N=202 for the three sites): "Did you find active involvement in the exercises helpful in stimulating your learning?"

47%	45%	7%	1%
Very Helpful	Some Help	Little Help	No Help

Alternate Approaches to Teacher Appraisal: Questionnaire Results

Ten major approaches to appraising teacher effectiveness were rated by workshop participants. They were asked to rate each approach according to the extent to which they would be willing to implement the approach in their school or school district. A brief listing of the various approaches is presented here. (For a complete description of each approach, see page 96.)

Approach

A. Unstructured supervisor observations

B. Systematic supervisor observations

C. Peer observations and ratings
D. Student observations and ratings
E. Teacher skill testing
F. Performance testing
G. Objectives-based testing results
H. Student growth based on results of nationally normed measures
I. Student growth based on results of locally normed measures
J. Student growth measured using regression techniques

Participants at all three workshop sites were unanimous in rating Approach H (Student performance on norm-referenced standardized tests) as the least desirable approach for appraising teacher effectiveness. Only 4% of the participants indicated they would be willing to implement this approach in their school(s).

The use of systematic supervisor observations (Approach B) was the most favored approach. However, there was no consensus on the approaches participants would be willing to implement. Only about half of the participants selected Approach B, the most popular choice. After systematic supervisory observations, the next most widely chosen approach was the use of objectives-based student achievement testing results for appraising teacher effectiveness (Approach G). These findings were interesting in that they were consistent with the results generally found with most CSE workshop audiences, although the NSPER participants were unique in their overall background as professional researchers.

The next most often selected choices were Approaches D and C, student observations ratings and peer observations and ratings. There was a great deal of variability in the selection of these approaches. For example, peer observations and ratings (Approach C) was the most frequently chosen approach at the Denver, Colorado workshop but finished in fourth place at the Atlanta, Georgia workshop.

In general, participants preferred implementing the process approaches (A, B, C, and D) over the product approaches (E, F, G, H, I, and J) in appraising teacher effectiveness. A notable

exception to this generalization was Approach G (Objectives-based testing results), a product oriented approach frequently selected. Finally, most participants indicated a willingness to utilize a combination of approaches in appraising teacher effectiveness rather than relying on a single technique.

ALTERNATE APPRAOCHES TO TEACHER APPRAISAL*

A. The principal and/or other supervisory personnel write a report of the teacher's effectiveness. This report is based on personal acquaintance with the teacher, comments by students, parents, and colleagues, and informal classroom observation.
B. The principal and/or other supervisory personnel systematically observe and rate the teacher using a rating form that lists important characteristics of good teaching. The teacher is given a score showing the extent to which he/she exhibits these characteristics in terms of a predetermined school or district standard of effectiveness.
C. The teacher's peers systematically observe and rate him/her on the extent to which he/she exhibits important characteristics of good teaching. The teacher must meet a predetermined school or district standard of effectiveness.
D. The teacher's students use a rating form to judge the extent to which the teacher exhibits important characteristics of good teaching. The teacher must meet a predetermined school or district standard of effectiveness.
E. The teacher is given written descriptions with an instructional objective, a sample test item measuring that objective, and information about the content it covers. A small group of students is randomly assigned to that teacher (to balance abilities), and is instructed by the teacher on the objective for one to ten lessons. After instruction, the students are tested on the objective. Teacher effectiveness is determined on the basis of how well the students achieved the objective.
G. The teacher instructs his/her students on relevant instructional objectives for a period of time ranging from several weeks to one semester. At the end of each instructional period, the students are tested on their attainment of those objectives. The level of student achievement provides the basis for

appraising teacher effectiveness.

H. Nationally or locally normed standardized tests are administered to students at the end of each school year. Teacher effectiveness is appraised on the basis of the extent to which student performance conforms to the test's established norms.

I. Nationally or locally normed standardized tests are administered to students to determine how much they increase their learning over time. The amount of desired increase is determined in advance by relevant school personnel, including teachers. Teacher effectiveness is appraised on the basis of the extent to which students achieve the desired increase.

J. The students in each grade or subject matter area are tested at the beginning and end of each school semester or year. Student scores are statistically analyzed to determine which class gains more or less than classes of comparable ability. Teacher appaisal is based on the proportion of "gainers" to "losers."

---

*Copyright 1973. Regents of the University of California.

	Very Helpful				Some Help				Little Help				No Help			
	D	S	A	T	D	S	A	T	D	S	A	T	D	S	A	T
Did you find active involvement in the exercises helpful in stimulating your learning?	25	23	46	94	33	18	40	91	7	4	3	14	0	1	1	2

Key: D = Denver
S = Seattle
A = Atlanta
T = Total

(N=between 196-203 participants from the three workshop sites answered each question.)

ALTERNATE APPROACHES TO TEACHER APPRAISAL: QUESTIONNAIRE

Rating Scale: (Willing to implement evaluation approach in school district)

1 = very unwilling
2 = not willing
3 = neutral
4 = willing
5 = very willing

	Approach	Denver Oct. 1974 AR*	Rank	Seattle Nov. 1974 AR	Rank	Atlanta Nov. 1974 AR	Rank
A.	Unstructured supervisor observations	2.66	6	2.22	2	2.81	2
B.	Systematic supervisor observations	3.14	10	3.05	9	3.29	8
C.	Peer observations and ratings	3.11	9	2.69	6	3.24	7
D.	Student observations and ratings	2.90	8	2.82	8	3.24	10
E.	Teacher skill testing	2.47	3	2.20	2	2.92	4
F.	Performance testing	2.72	7	2.60	5	3.17	6
G.	Objectives-based testing results	2.00	1.5	3.10	10	3.36	9
H.	Student growth based on results of nationally normed measures	2.00	1.5	1.86	1	2.13	1

101

I.	Student growth based on results of locally normed measures	2.60	4	2.76	7	2.82	3
J.	Student growth measured using regression techniques	2.61	5	2.58	4	3.11	5

Rank: 10 = highest
1 = lowest

*AR = Average of Ratings

OCCASIONAL PAPERS

1. THE PROBLEM AND PROBLEM DELINEATION TECHNIQUES – William J. Gephart, Phi Delta Kappa. Presented at the Second National Symposium for Professors of Educational Research, sponsored by Phi Delta Kappa, Boulder, Colorado, November 21, 1968. A discussion of the nature of the concept "problem" as related to educational research with a discussion of several techniques useful in problem identification and delineation. $1.00
2. A REVIEW OF INSTRUMENTS DEVELOPED TO BE USED IN THE EVALUATION OF THE ADEQUACY OF REPORTED RESEARCH – Bruce B. Bartos, Phi Delta Kappa & Indiana University. Presented at the Annual Meeting of the American Educational Research Association, February, 1969, Los Angeles, California. A brief description and bibliographic annotation of 40 instruments developed to be used in assessing the methodological quality of completed research. $.25
3. PROFILING EDUCATIONAL RESEARCH – William J. Gephart, Phi Delta Kappa, January, 1969. The rationale for the development of a methodology profile on completed research to show its strengths and weaknesses. Included are flow charts for profiling the five facets of the research process. $.75
4. APPLICATION OF THE CONVERGENCE TECHNIQUE TO READING – William J. Gephart, Phi Delta Kappa, January, 1969. An interim report on a research program planning effort in the field of reading. (Out of Print)
5. THE CONVERGENCE TECHNIQUE AND READING: A PROGRESS REPORT – William J. Gephart, Phi Delta Kappa. Presented at the Annual Meeting of the International Reading Association, May 2, 1969, Kansas City, Missouri. A second interim report on the planning of a reading research program. (Out of Print)
6. THE EIGHT GENERAL RESEARCH METHODOLOGIES: A FACET ANALYSIS OF THE RESEARCH PROCESS – William J. Gephart, Phi Delta Kappa, July 14, 1969. The identification and description of general research methods in education through the use of Gutman's facet design and analysis technique. It also details the procedures for the Gutman technique. This paper was printed in the proceedings of the Warsaw, Poland Congress of the International Association for the Advancement of Educational Research. $1.00
7. PROFILING INSTRUCTIONAL PACKAGE – William J. Gephart & Bruce B. Bartos, Phi Delta Kappa, August, 1969. An instruction text to assist individuals with no prior research training in the use of research profiling flow charts to assess the methodological adequacy of completed research. $1.00
8. EDUCATIONAL KNOWLEDGE USE – Gene V. Glass, Laboratory of Educational Research, University of Colorado. An analysis of the availability and use of empirically based information in education. $.50
9. MEASUREMENT AND RESEARCH IN THE SERVICE OF EDUCATION – Warren G. Findley, Research and Development Center in Educational Stimulation, University of Georgia. Originally presented as an invited address at the annual meeting of the American Educational Research Association; this paper uses an historical perspective to examine the role of measurement and research in education. $.75
10. THE EDUCATIONAL CATALYST: AN IMPERATIVE FOR TODAY – Joe H. Ward, Jr., Reeve Love, and George M. Higginson, Southwest Educational Development Laboratory, Austin, Texas, July, 1971. An analysis of the problems involved in the process of change and improvement of the practice of education. This paper poses a new professional speciality for the facilitation of empirically based educational improvements. $1.00

11. DISSERTATIONS YOU MAY WANT TO SEE – William J. Gephart, Phi Delta Kappa, 1970. A collection of dissertations done in 1969 which focus on research training. $.25

12. THE DOCTORATE IN EDUCATION IN CANADA – Neville L. Robertson, Commission on Higher Education, Phi Delta Kappa, 1971. An analysis of the institutions offering the doctorate in education in Canada. This paper is a companion piece to a larger study of similar institutions in the United States. $.75

13. THE IMPORTANCE OF STATISTICAL POWER IN EDUCATIONAL RESEARCH – John K. Miller, and Thomas R. Knapp, University of Rochester. When an educational experiment results in non-significant differences can it be said that no difference exists? This paper discusses the concept that must be attended to IF that question is to be answered. It also details the procedure for determination of sample size needed in an experiment. $1.25

14. THE CONDUCT OF DEVELOPMENT IN EDUCATION – Richard E. Schutz, SWRL Educational Research and Development. The past decade has developed awareness of the tools and procedures needed to perform educational tasks. Some have argued that the research process is the vehicle for achieving that end. Schutz argues that the development process is different from the research process in many important ways that need to be understood by educators. $1.50

15. THE PROCESS OF DEVELOPMENT – Gerald Nadler and William J. Gephart. The development process is examined from five points of view in this interview of Gerald Nadler, Chairman of the Industrial Engineering Department, University of Wisconsin. Questions posed by Gephart and their discussion examine the general nature of the process, the types of problems for which it is an appropriate resolution strategy, the component activities of the development process, the types of products resulting from it, and the criteria that would be used in assessing the quality of a development effort. $1.50

16. THE MOST SIGNIFICANT EDUCATIONAL RESEARCH CONTRIBUTIONS OF THE PAST TEN YEARS – William J. Gephart, Phi Delta Kappa, in collaboration with Berlie Fallon, Anton J. Netusil, and William K. Poston, Jr. Three individuals were asked to identify important research contributions. Work thus identified is briefly summarized and comments are made about the manner in which educational research *ought* to be looked to by educators. $1.25

17. EVALUATION: PAST, PRESENT, AND FUTURE – William J. Gephart, Phi Delta Kappa. Evaluation's history is traced from the movements in measurement, in the Eight-Year Study, and in accreditation. Next evaluation is examined as it has grown from 1965 to 1975. A schema is presented for integrating the many, many models of evaluation that have come into being during that decade. Finally, the paper suggests some possible directions for evaluation's future given current focus in the field. $1.50.